

Mathématiques Assistées par Ordinateur -
Probabilités et Statistiques

Paul Melotti
Basé sur des notes de cours de Yan Pautrat

Master 1
2021-2022

version du 13 juin 2023

Table des matières

0	Rappels et commandes Python	5
0.1	L'aléatoire en Python	5
0.2	Illustration de données	6
0.3	Lois classiques	6
0.3.1	Lois discrètes	7
0.3.2	Lois continues	8
1	Simulation de variables aléatoires	9
1.1	Lois discrètes	9
1.2	Méthode par inversion	10
1.3	Méthode de rejet	12
1.4	Conditionnement	14
1.5	Exercices	15
1.6	Générateurs pseudo-aléatoires (facultatif)	16
2	Convergence des variables aléatoires	19
2.1	Rappels	19
2.2	Illustration de la convergence presque-sûre	22
2.3	Illustration de la convergence en loi	23
2.4	Illustration de la convergence \mathbb{P}	25
2.5	Illustration des convergences L^p	25
2.6	Exercices	26
3	Grands théorèmes de convergence	29
3.1	Lois des grands nombres	29
3.2	Théorèmes centraux limite	32
3.3	Valeurs extrêmes	35
3.4	Principes de grandes déviations	36
4	Tests et estimateurs classiques	39
4.1	Estimateurs	39
4.1.1	Définitions	39
4.1.2	Méthode des moments	41
4.1.3	Méthode par insertion	41
4.1.4	Méthode du maximum de vraisemblance	41
4.2	Borne de Cramér-Rao et modèles exponentiels	42

4.2.1	Minoration du risque	42
4.3	Intervalles de confiance	43
4.4	Tests d'hypothèses : définitions générales	45
4.5	Tests du chi-deux	47
4.5.1	Ajustement à une loi	47
4.5.2	Ajustement à une famille de lois	48
4.6	Test de Kolmogorov et dérivés	49
4.7	Exercice supplémentaire	50
5	Chaînes de Markov	53
5.1	Simulation et résultats classiques	53
5.1.1	Trajectoire	53
5.1.2	Irréductibilité	54
5.1.3	Période	54
5.1.4	Mesure invariante et théorème ergodique	54
5.1.5	Convergence en loi vers l'équilibre	55
5.2	Méthodes de Monte-Carlo	56
5.3	Algorithme de Metropolis–Hastings	57
5.4	Mesures de Gibbs	59
5.5	Méthode du recuit simulé	61
5.5.1	Algorithme du recuit	61
5.5.2	Vitesse de convergence : méthode spectrale	64
5.6	Exercice supplémentaires	64

Chapitre 0

Rappels et commandes Python

Commençons par donner les bases de la simulation en Python, et de la représentation de données.

0.1 L'aléatoire en Python

Pour des rappels généraux sur le fonctionnement de Python, vous pouvez consulter en préambule le polycopié de Sophie Lemaire, que vous avez peut-être déjà pratiqué :

<http://www.math.u-psud.fr/~lemaire/poly13python.pdf>

Vous êtes également encouragé-e-s à vous servir de l'aide de Python, soit en ligne soit directement depuis votre gestionnaire, quel qu'il soit. On commencera toujours par invoquer les modules suivants :

```
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import numpy.random as rnd
import scipy.stats as sts
```

Vous connaissez sans doute déjà les bibliothèques `numpy` et `scipy`, qui contiennent un grand nombre de fonctions mathématiques et de structures de données utiles. Vous connaissez certainement celle que l'on a nommée `plt`, qui permet de faire des graphiques.

Le paquet `random` de `numpy` que nous avons abrégé en `rnd`, permet de réaliser des simulations indépendantes de lois classiques. La syntaxe est particulièrement simple. Allez voir la page en ligne pour quelques exemples.

Le paquet `stats` de `scipy` que nous avons abrégé en `sts` dans notre préambule, permet de réaliser des simulations indépendantes de lois classiques, mais aussi d'avoir accès aux densités, fonctions de répartition, quantiles... de ces lois. La philosophie est un peu plus « orientée objet » : par exemple, si on travaille avec une variable aléatoire X de loi $\mathcal{N}(1,2^2)$, taper `X=sts.norm(1,2)` ne renvoie pas une réalisation mais un objet qui représente la variable aléatoire dans sa globalité. Pour avoir une simulation aléatoire, on pourra taper `X.rvs()`. Pour connaître la densité de cette loi en $x = 5$ on peut faire appel directement à

`X.pdf(5)` (ou à `sts.norm.pdf(5,1,2)` mais il faut alors faire attention à l'ordre des variables).

0.2 Illustration de données

Les principales commandes pour illustrer des données sont les suivantes ; elles sont toutes issues du paquet `matplotlib.pyplot`, que nous avons abrégé en `plt`. N'hésitez pas à consulter l'aide quand vous voulez les utiliser :

- `plt.plot` pour les tracés de courbes,
- `plt.step` pour les tracés de fonctions en escalier,
- `plt.stem` pour les diagrammes bâton,
- `plt.scatter` pour les nuages de points,
- `plt.hist` pour les histogrammes.

Pour les quatre premières, la syntaxe de base est `plt.commande(X,Y)` où `X` et `Y` sont des listes ou tableaux de réels et de même longueur. La syntaxe de `plt.hist` est, forcément différente : `plt.hist(X)` trace un histogramme obtenu en regroupant les valeurs contenues dans `X` dans des classes, la hauteur étant proportionnelle au nombre de points tombant dans la classe. On peut/doit utiliser les options suivantes :

- `bins=c` (où `c` est un entier) imposera `c` classes, `bins='auto'` fait un choix automatique ;
- `density=True` normalise les hauteurs pour avoir une surface totale égale à 1, `density=False` conserve une hauteur égale au nombre de points.

Exercice 1 *On se demande à présent comment tracer de manière la fonction de répartition empirique \widehat{F}_N d'un N -échantillon $Y^{(1)}, \dots, Y^{(N)}$. Celle-ci est définie par*

$$\widehat{F}_N(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \leq x}.$$

On note $Y_{(1)}, \dots, Y_{(N)}$ la statistique d'ordre associée à $Y^{(1)}, \dots, Y^{(N)}$ (autrement dit, les $Y_{(1)}, \dots, Y_{(N)}$ sont les $Y^{(1)}, \dots, Y^{(N)}$ ordonnés par ordre croissant). Si `Y` est un vecteur contenant les valeurs $Y^{(1)}, \dots, Y^{(N)}$, comment obtenir un vecteur `Yord` contenant les $Y_{(i)}$ (dans le bon ordre) ? Combien vaut \widehat{F}_N sur l'intervalle $[Y_{(i)}, Y_{(i+1)})$? Comment tracer une fonction en escalier ?

De manière générale, une loi discrète (à support suffisamment petit) sera représentée par un diagramme à bâtons ; pour une loi réelle à densité, on pourra envisager un histogramme (mais on a peu de contrôle sur celui-ci a priori) ; dans le doute, la fonction de répartition empirique sera souvent préférée, car elle couvre tous les cas.

0.3 Lois classiques

On rappelle ici les lois les plus classiques, et leurs notations qui seront utilisées tout au long de ces notes.

0.3.1 Lois discrètes

Mesure de Dirac δ_x : pour $x \in \mathbb{R}^d$, c'est une loi qui vaut x avec probabilité 1.

Loi de Bernoulli $\mathcal{B}(p)$ de paramètre $p \in [0,1]$: elle vaut 1 avec probabilité p et 0 avec probabilité $1 - p$. Autrement dit

$$\mathcal{B}(p) = p\delta_1 + (1 - p)\delta_0.$$

Loi binomiale $\mathcal{B}(n,p)$ de paramètres $n \in \mathbb{N}^*$ et $p \in [0,1]$: elle est à support dans $\{0, \dots, n\}$ et vaut $k \in \{0, \dots, n\}$ avec probabilité

$$\binom{n}{k} p^k (1 - p)^{n-k}.$$

Une variable aléatoire de loi $\mathcal{B}(n,p)$ est égale en loi à la somme de n variables aléatoires indépendantes de loi $\mathcal{B}(p)$. Ainsi, dans le cas $n = 1$, c'est une généralisation du cas précédent.

Loi multinomiale de paramètres $n \in \mathbb{N}^*$ et $p_1, \dots, p_m \in [0,1]$ tels que $\sum_{i=1}^m p_i = 1$: si on effectue n tirages indépendants, chacun étant à valeurs dans $\{1, \dots, m\}$ et de loi (p_1, \dots, p_m) , la loi du nombre de chaque valeur obtenue (N_1, \dots, N_m) est multinomiale de paramètres n, p_1, \dots, p_m . Pour tous $(n_1, \dots, n_m) \in \mathbb{N}^m$ tels que $\sum_{i=1}^m n_i = n$,

$$\mathbb{P}(N_1 = n_1, \dots, N_m = n_m) = \frac{n!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m}.$$

Loi uniforme discrète $\mathcal{U}(\{a_1, \dots, a_n\})$: pour un ensemble fini $\{a_1, \dots, a_n\}$, la loi uniforme discrète est définie comme

$$\mathcal{U}(\{a_1, \dots, a_n\}) = \frac{1}{n} \sum_{k=1}^n \delta_{a_k}.$$

Autrement dit, tous les éléments sont équiprobables de probabilité $\frac{1}{n}$.

Loi géométrique $\mathcal{G}(p)$ de paramètre $p \in]0,1]$: c'est la loi du moment du premier succès dans un jeu de pile ou face avec probabilité p de gagner et $q = 1 - p$ de perdre. Dans la convention la plus courante, est portée par \mathbb{N}^* , et la probabilité de $k \in \mathbb{N}^*$ est pq^{k-1} ; on trouve aussi la convention où elle est portée par \mathbb{N} , et la probabilité de $k \in \mathbb{N}$ est pq^k . Pour distinguer les deux, on écrit parfois $\mathcal{G}_{\mathbb{N}^*}(p)$ et $\mathcal{G}_{\mathbb{N}}(p)$.

Loi binomiale négative $\text{BinNeg}(n,p)$ de paramètres $n \in \mathbb{N}^*, p \in [0,1]$: elle généralise la loi géométrique sur \mathbb{N} , c'est la loi du nombre d'échecs avant l'obtention du n -ième succès. Ainsi, pour tout $k \in \mathbb{N}$, la probabilité d'obtenir k est

$$\binom{k+n-1}{k} p^n (1-p)^k$$

Loi hypergéométrique $\mathcal{H}(n,p,N)$ de paramètres $n \in \mathbb{N}^*, p \in [0,1], N \geq n$: on tire simultanément n boules dans une urne contenant N boules parmi lesquelles pN sont rouges et $(1-p)N$ sont bleues. Le nombre X de boules rouges tirées suit une loi $\mathcal{H}(n,p,N)$, caractérisée par

$$\mathbb{P}(X = k) = \frac{\binom{pN}{k} \binom{(1-p)N}{n-k}}{\binom{N}{n}}.$$

Loi de Poisson $\mathcal{P}(\lambda)$ de paramètre $\lambda > 0$: elle est portée par \mathbb{N} , et la probabilité de $k \in \mathbb{N}$ est

$$e^{-\lambda} \frac{\lambda^k}{k!}.$$

0.3.2 Lois continues

Loi uniforme $\mathcal{U}([a,b])$ sur l'intervalle $[a,b]$ avec $a < b$: c'est la loi de densité

$$\frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

Loi exponentielle $\mathcal{E}(\lambda)$ de paramètre $\lambda > 0$: c'est la loi de densité

$$\lambda \exp(-\lambda x) \mathbb{1}_{\mathbb{R}_+}(x).$$

Loi normale (ou gaussienne) $\mathcal{N}(m,\sigma^2)$ de moyenne $m \in \mathbb{R}$ et de variance $\sigma^2 > 0$: c'est la loi de densité

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

Loi de Cauchy $\mathcal{C}(\lambda)$ de paramètre $\lambda > 0$: c'est la loi de densité

$$\frac{1}{\pi} \frac{\lambda}{\lambda^2 + x^2}.$$

Cette loi n'est pas dans L^1 : une variable aléatoire de Cauchy n'admet pas d'espérance.

Loi Gamma $\Gamma(a,\lambda)$ de paramètres $a > 0$ et $\lambda > 0$: c'est la loi de densité

$$\frac{\lambda^a}{\Gamma(a)} x^{a-1} \exp(-\lambda x) \mathbb{1}_{\mathbb{R}_+}(x)$$

où dans cette formule, $\Gamma(a)$ désigne la valeur en a de la fonction Gamma d'Euler.

On trouve parfois une autre convention pour les paramètres, avec $\theta = \lambda^{-1}$. Dans le module `scipy.stats`, le paramètre ici noté α s'appelle `a` et celui noté $\theta = \lambda^{-1}$ s'appelle `scale`.

Loi du khi-deux $\chi^2(k)$ à $k \in \mathbb{N}^*$ degrés de liberté : c'est la loi de la somme des carrés de k variables iid $\mathcal{N}(0,1)$. Elle a pour densité

$$\frac{1}{2^{k/2} \Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \mathbb{1}_{\mathbb{R}_+}(x).$$

Rappelons aussi qu'il existe des mesures de probabilité sur \mathbb{R} qui ne sont ni discrètes, ni à densité par rapport à la mesure de Lebesgue.

Chapitre 1

Simulation de variables aléatoires

Dans ce chapitre, on se donne une suite de variables i.i.d. (U_1, U_2, \dots) uniformément distribuées dans $[0,1]$ (on peut penser à des appels successifs à la fonction `random()` du paquet `numpy.random` par exemple). On veut l'utiliser pour simuler n'importe quelle autre variable aléatoire de notre choix.

Dans la plupart des cas, on pourrait atteindre notre objectif en utilisant d'autres fonctions des paquets `numpy.random` ou `scipy.stats`. Mais l'objectif est de comprendre comment ces commandes fonctionnent, d'acquérir des bases théoriques, et de savoir simuler des variables plus exotiques si on en rencontre un jour.

Bien sûr cela repousse le problème : comment l'ordinateur produit-il des nombres aléatoires uniformes dans $[0,1]$? Pour les plus curieux, on donne quelques éléments de réponse à la fin de ce chapitre.

1.1 Lois discrètes

Commençons par un exercice instructif :

Exercice 2 Soit $p \in]0,1[$. Écrire une fonction qui simule une loi de Bernouilli $\mathcal{B}(p)$ à l'aide de la fonction `random`.

Une loi discrète est à support dans un ensemble au plus dénombrable. Si on a compris le cas de la variable de Bernouilli, le résultat suivant n'est qu'une généralisation.

Proposition 1.1.1 (Simulation canonique de lois discrètes)

Soit $\mu = \sum_{k \in \mathbb{N}} p_k \delta_{x_k}$ une loi sur un ensemble au plus dénombrable $\{x_0, x_1, \dots\}$. Soit $s_{-1} = 0$ et pour tout $k \in \mathbb{N}$, $s_k = s_{k-1} + p_k$. Soit U une variable aléatoire uniforme sur $[0,1]$. Alors la variable aléatoire

$$X = \sum_{k \in \mathbb{N}} x_k \mathbb{1}_{[s_{k-1}, s_k[}(U)$$

suit la loi μ .

Exercice 3 Prouver la Proposition 1.1.1.

Exercice 4 Écrire un programme qui prend en entrée un vecteur de probabilités **probas** et retourne une réalisation d'une variable aléatoire à valeurs dans $\{0, \dots, n-1\}$ et de loi donnée par le vecteur de probabilités **probas**.

Le programme suggéré par cet exercice ne permet pas *a priori* de simuler une loi dont le support est infini dénombrable, comme une loi géométrique par exemple. De plus, pour certaines lois spécifiques, il peut exister des méthodes plus efficaces. On va en voir quelques unes.

Proposition 1.1.2 (Simulation de la loi uniforme discrète) Soit $n \in \mathbb{N}^*$. Si U est une variable aléatoire uniforme sur $[0,1]$ alors la variable aléatoire $\lfloor nU \rfloor + 1$ suit la loi uniforme discrète sur $\{1, \dots, n\}$.

Proposition 1.1.3 (Simulation de la loi géométrique) Soit $p \in]0,1[$. Si U est une variable aléatoire uniforme sur $[0,1]$ alors la variable aléatoire $\lfloor \frac{\log U}{\log(1-p)} \rfloor + 1$ suit la loi géométrique $\mathcal{G}(p)$.

Exercice 5 Prouver les Propositions 1.1.2 et 1.1.3. On pourra au choix montrer qu'il s'agit en fait d'applications de la Proposition 1.1.1, ou utiliser une méthode directe.

Contrairement aux exemples précédents, la loi de Poisson n'est pas facile à simuler par la méthode canonique. On pourra tout de même la simuler par une méthode *ad hoc* :

Proposition 1.1.4 (Simulation de la loi de Poisson) Soit $\lambda > 0$. Si $(U_n)_{n \in \mathbb{N}^*}$ est une suite i.i.d. de variables aléatoires uniformes sur $[0,1]$ alors la variable aléatoire

$$N = \inf\{n \geq 0 \mid U_1 \times \dots \times U_{n+1} < e^{-\lambda}\}$$

est finie presque sûrement et suit la loi de Poisson $\mathcal{P}(\lambda)$.

1.2 Méthode par inversion

On va maintenant passer à des lois μ réelles quelconques. L'objet fondamental de cette section est la fonction de répartition, définie par $F(x) = \mu(-\infty, x]$. On rappelle que F est une fonction croissante, en tout point continue à droite avec une limite à gauche, a une quantité au plus dénombrable de points de discontinuité, et que $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow +\infty} F(x) = 1$. On définit son pseudo-inverse $F^{(-1)}$ comme la fonction

$$F^{(-1)} :]0,1[\rightarrow \mathbb{R} \\ q \mapsto \inf\{x \in \mathbb{R} \mid F(x) \geq q\}.$$

Le résultat suivant dit que si l'on sait simuler la loi uniforme sur $[0,1]$, alors on sait simuler la loi μ :

Théorème 1.2.1 Soit μ une mesure de probabilité sur \mathbb{R} , F sa fonction de répartition et $F^{(-1)}$ son pseudo-inverse. Alors si U est une variable de loi uniforme sur $[0,1]$, la variable $F^{(-1)}(U)$ a pour loi μ .

Exercice 6

1. Montrer que $q \leq F(x) \Leftrightarrow F^{(-1)}(q) \leq x$. On pourra montrer que l'ensemble $I_q = \{x \in \mathbb{R} \mid F(x) \geq q\}$ est un intervalle, minoré et non majoré, fermé à gauche.
2. Soit $U \sim \mathcal{U}([0,1])$. En calculant sa fonction de répartition, montrer que $F^{(-1)}(U)$ suit la loi μ .

On a donc une méthode très générale et directement applicable de simulation de la loi μ sur \mathbb{R} : on calcule (à la main) $F^{(-1)}$, on simule une uniforme U et on choisit $X = F^{(-1)}(U)$.

Remarque 1.2.2 Dans le cas où μ est une loi discrète, le Théorème 1.2.1 se ramène en fait à la Proposition 1.1.1.

L'exercice suivant propose des applications pratiques.

Exercice 7 Appliquez la méthode du Théorème 1.2.1 pour écrire des fonctions permettant de simuler des échantillon de :

1. géométrique $\mathcal{G}(p)$ où $p \in]0,1[$,
2. exponentielle $\mathcal{E}(\lambda)$ où $\lambda > 0$,
3. de Cauchy $\mathcal{C}(\lambda)$,
4. de Weibull¹ $\mathcal{W}(a,\lambda)$.

Tenter, et échouer, de calculer $F^{(-1)}$ dans le cas où μ est une loi normale.

Cela montre en particulier que $\frac{-\log U}{\lambda}$ suit la loi $\mathcal{E}(\lambda)$, qui est un résultat assez utile.

Cette méthode très générale a cependant plusieurs défauts pratiques, comme on le voit avec de cas le la loi normale. Dans les paragraphes qui suivent on va donner d'autres méthodes, qui permettront notamment de simuler la loi normale.

Avant cela, donnons quand même un résultat « théorique » qu'on peut voir comme une conséquence de la méthode d'inversion.

Théorème 1.2.3 (Représentation de Skorokhod) Soit $(X_n)_{n \geq 0}, X$ des v.a. à valeurs dans \mathbb{R}^d , telles que X_n converge en loi vers X . Alors il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et des v.a. $(Y_n)_{n \geq 0}, Y$ définies sur Ω telles que

- $\forall n \geq 0, Y_n$ a même loi que X_n ,
- Y a même loi que X ,
- Y_n converge p.s. vers Y .

1. qui pour $a > 0$ et $\lambda > 0$ a pour densité $f(x) = a\lambda x^{a-1} \exp(-\lambda x^a) \mathbb{1}_{\mathbb{R}_+}(x)$.

Ce théorème est à prendre avec des pincettes, en particulier remarquez que la loi jointe de (Y_n, Y_m) n'a aucune raison d'être la même que celle de (X_n, X_m) .

Preuve. [Idée de démonstration] Dans le cas $d = 1$, prenons pour espace de probabilité $]0,1[$ muni de la tribu borélienne et de la mesure de Lebesgue. On pose F_n , resp. F , la fonction de répartition de X_n , resp. X . On pose alors, pour $\omega \in \Omega$,

$$\forall n \geq 0, Y_n(\omega) = F_n^{(-1)}(\omega), \text{ et } Y(\omega) = F^{(-1)}(\omega),$$

On sait que pour tout t point de continuité de F , $F_n(t) \rightarrow_{n \rightarrow \infty} F(t)$. Cela permet de montrer, avec un peu d'analyse (voir la page wikipedia) que $Y_n(\omega) \rightarrow Y(\omega)$ pour tout $\omega \notin C$, où C est l'ensemble des « valeurs plateau » de F . Formellement, $C = \{q \in]0,1[\mid \exists x \neq y, F(x) = F(y) = q\}$, et on peut montrer que cet ensemble est au plus dénombrable, ce qui démontre bien la convergence p.s.

Pour passer à la dimension d , voir le chapitre qui vient sur le conditionnement.

1.3 Méthode de rejet

Une autre méthode générale est la méthode de rejet, qui permet de simuler une loi μ conditionnée à prendre des valeurs dans une sous-partie B de \mathbb{R}^d , dès que l'on sait simuler la loi μ non conditionnée. Comme on va le voir, une application directe mais particulièrement intéressante est la simulation d'une loi μ à densité f , dès que l'on sait simuler une loi ν à densité g , avec $f \leq \lambda g$ pour un certain $\lambda > 0$.

Proposition 1.3.1 *Soit $(X_n)_n$ une suite de variables aléatoires indépendantes de loi μ et B un borélien tel que $\mu(B) > 0$. Soit T le plus petit entier $n \geq 1$ tel que $X_n \in B$. Alors*

1. T est une variable aléatoire de loi géométrique de paramètre $\mu(B)$,
2. X_T est une variable aléatoire indépendante de T ayant pour loi la loi conditionnelle $\mu(\cdot|B) := \frac{\mu(\cdot \cap B)}{\mu(B)}$.

Preuve. Pour A borélien de \mathbb{R}^d et $n \in \mathbb{N}^*$,

$$\begin{aligned} \mathbb{P}(T = n, X_T \in A) &= \mathbb{P}(X_1 \notin B, \dots, X_{n-1} \notin B, X_n \in A \cap B) \\ &= (1 - \mu(B))^{n-1} \mu(A \cap B) \\ &= (1 - \mu(B))^{n-1} \mu(B) \mu(A|B) \end{aligned}$$

qui est bien la probabilité pour la loi jointe annoncée. \square

Une application directe de la Proposition 1.3.1 est le tirage uniforme suivant la loi uniforme sur un Borélien B de \mathbb{R}^d , quand on sait effectuer un tirage uniforme sur un autre Borélien $C \supset B$. En voici un exemple :

Exercice 8 *Écrire une fonction Python tirant un point uniformément dans le disque unité D . On pourra utiliser une instruction `while`. En moyenne, combien de tirages de l'uniforme sur $[-1, +1]^2$ faut-il pour obtenir un tirage de l'uniforme sur D ?*

Cette méthode permet également de simuler des lois à densité. On utilisera pour cela les deux résultats suivants, parfois appelés théorèmes densité-surface.

Proposition 1.3.2 *Soit ν une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ admettant une densité g par rapport à la mesure de Lebesgue. Soit X_ν une variable aléatoire de loi ν et U une variable aléatoire indépendante de X_ν , de loi $\mathcal{U}([0,1])$. Alors la variable $(X_\nu, Ug(X_\nu))$ suit une loi uniforme sur*

$$E_g = \{(x,y) \in \mathbb{R}^d \times \mathbb{R}_+ \mid 0 < y < g(x)\}.$$

Preuve. Soit $\phi : \mathbb{R}^d \times \mathbb{R}^+ \rightarrow \mathbb{R}$ mesurable bornée, alors

$$\begin{aligned} \mathbb{E}[\phi(X, Ug(X))] &= \int_{\mathbb{R}^d \times]0,1[} \phi(x, ug(x))g(x)dxdu \\ &= \int_{\mathbb{R}^d} \left(\int_0^1 \phi(x, ug(x))g(x)du \right) dx \\ &= \int_{\mathbb{R}^d} \left(\int_0^{g(x)} \phi(x, v)dv \right) dx \\ &= \int_{E_g} \phi(x, v)dx dv. \end{aligned}$$

On a également un résultat réciproque, dont la preuve est similaire :

Lemme 1.3.3 *Soit (X, Y) une variable aléatoire de loi uniforme sur E_g comme ci-dessus. Alors X suit la loi de densité g par rapport à la mesure de Lebesgue.*

Une conséquence immédiate des Propositions 1.3.1 et 1.3.2 est la suivante :

Proposition 1.3.4 *Soient μ et ν deux mesures de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ de densités f, g respectivement par rapport à la mesure de Lebesgue, et telles que $f \leq \lambda g$ Lebesgue-presque partout, pour un certain $\lambda > 0$. Soit $(X_n)_n$ une suite de variables aléatoires indépendantes de loi ν , $(U_n)_n$ une suite de variables aléatoires indépendantes de loi uniforme $\mathcal{U}([0,1])$, et T le plus petit entier $n \geq 1$ tel que $\lambda U_n g(X_n) \leq f(X_n)$. Alors*

1. T est une variable aléatoire de loi géométrique de paramètre $1/\lambda$,
2. X_T est une variable aléatoire de loi μ .

Preuve. La Proposition 1.3.2 montre que $(X_n, U_n \lambda g(X_n))$ est une suite i.i.d. de loi uniforme sur $E_{\lambda g}$. La Proposition 1.3.1 montre à son tour que T suit une loi géométrique de paramètre $\frac{\int f(x) dx}{\int \lambda g(x) dx} = 1/\lambda$, et que $(X_T, U_T \lambda g(X_T))$ suit une loi uniforme sur E_f . Le Lemme 1.3.3 montre que X_T suit une loi de densité f par rapport à la mesure de Lebesgue. \square

La Proposition 1.3.4 donne à nouveau une méthode générale et facilement applicable. Le cas le plus simple est celui d'une densité f sur un intervalle borné

de \mathbb{R} , acceptant une borne uniforme, auquel cas on peut choisir g constante. Remarquez que l'on a intérêt à choisir le λ le plus petit possible (mais respectant, bien sûr, la contrainte $f \leq \lambda g$), puisque $\mathbb{E}[T] = \lambda$.

Exercice 9 *On cherche à simuler la loi normale centrée réduite $\mathcal{N}(0,1)$.*

1. *Montrer que si X a pour loi $\mathcal{N}(0,1)$, alors $|X|$ suit la loi de densité*

$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \mathbb{1}_{\mathbb{R}_+}(x).$$

Cette loi s'appelle la loi demi-normale.

2. *On considère une variable Y qui suit la loi $\mathcal{E}(1)$. Rappeler sa densité g et trouver un $\lambda > 0$ tel que $f \leq \lambda g$.*
3. *En déduire un algorithme permettant de simuler la loi demi-normale par rejet.*
4. *Si on sait simuler une loi demi-normale, comment simuler une loi normale ?*

Remarque 1.3.5 *Une manière de choisir une loi ν facilement simulable et donnant un λ petit est de polygonaliser le domaine E_f et d'utiliser l'algorithme décrit dans l'exercice 14.*

1.4 Conditionnement

Revenons sur la notion de loi conditionnelle, qui a un sens assez naturel du point de vue de la simulation. Pour tout couple de variable aléatoires (X,Y) , on peut décider de simuler d'abord Y , puis de simuler X à partir de ce résultat. Pour formaliser cela, on a besoin de la famille des lois conditionnelles « de X sachant $Y = y$ » pour toutes les valeurs possibles de y . Le principe général est :

Si l'on sait simuler la loi de Y et toutes les lois conditionnelles de X sachant $Y = y$, alors on sait simuler la loi de (X,Y) .

Supposons que X est à valeurs dans \mathbb{R}^n , Y à valeurs dans \mathbb{R}^m , et ν de loi ν . Pour simuler X à partir de Y , on souhaite trouver une famille de mesures de probabilités sur \mathbb{R}^n notées $(\mu_y)_{y \in \mathbb{R}^m}$ telles que pour ϕ mesurable bornée,

$$\mathbb{E}[\phi(X,Y)] = \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^n} \phi(x,y) \mu_y(dx) \right) \nu(dy).$$

De notre point de vue, on tire Y selon la loi ν , et on tire X selon la loi conditionnelle relativement à $\{Y = y\}$, notée μ_y . Le théorème de Jirina affirme qu'une famille de lois conditionnelles existe toujours (avec un résultat de régularité : pour tout borélien A de \mathbb{R}^n , la fonction $y \mapsto \mu_y(A)$ est mesurable).

Entre autres, cela permet de définir l'espérance conditionnelle :

$$\mathbb{E}[\phi(X) | Y] = h(Y) \text{ où } h(y) = \int_{\mathbb{R}^n} \phi(x) \mu_y(dx).$$

Exemple 1 Si le couple (X, Y) a pour densité $h(x, y)$, alors il est facile de voir que Y a pour densité $g(y) = \int_{\mathbb{R}^n} h(x, y) dx$, puis de trouver la famille de lois conditionnelles qui fait marcher (1.4) :

$$\mathbb{E}[\phi(X, Y)] = \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^n} \phi(x, y) \frac{h(x, y)}{g(y)} dx \right) g(y) dy.$$

On retrouve que la loi de X conditionnée à $Y = y$ est à densité $\mu_y(dy) = \frac{h(x, y)}{g(y)} dy$ (stricto sensu, pour les y tels que $g(y) > 0$).

Exercice 10 (*) En utilisant le conditionnement et le Théorème 1.2.1, démontrer le résultat suivant :

Pour toute loi μ sur \mathbb{R}^d , il existe une fonction borélienne $f_\mu : \mathbb{R}^d \rightarrow \mathbb{R}^d$ dont l'ensemble des points de discontinuité est Lebesgue-négligeable et telle que, si U_1, \dots, U_d sont des variables i.i.d uniformes sur $[0, 1]$, la variable $f_\mu(U_1, \dots, U_d)$ suit la loi μ .

Ce résultat justifie qu'on puisse simuler « tout ce que l'on veut » à l'aide de variables uniformes indépendantes, mais en pratique l'explicitation d'une fonction f_μ n'est pas forcément simple, ni même possible.

1.5 Exercices

Exercice 11 Si B_1, \dots, B_n sont des variables de Bernoulli indépendantes de loi $\mathcal{B}(p)$, la variable $B_1 + \dots + B_n$ suit une loi binomiale $\mathcal{B}(n, p)$. Utiliser ce résultat pour donner une fonction qui simule une loi binomiale de paramètres n et p .

Exercice 12 (Algorithme de Box-Müller) On donne ici la méthode classique de simulation des lois normales. Soient U_1 et U_2 sont deux variables aléatoires indépendantes de loi $\mathcal{U}([0, 1])$. Montrer que les variables X_1 et X_2 définies par

$$X_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2) \quad X_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2)$$

sont indépendantes et de même loi $\mathcal{N}(0, 1)$. En déduire une fonction qui simule un échantillon (X_1, \dots, X_n) de variables aléatoires i.i.d. de loi $\mathcal{N}(0, 1)$.

Exercice 13 Montrez la réciproque partielle suivante de l'exercice 6 : si X une variable aléatoire de fonction de répartition F continue, montrez que $F(X) \sim \mathcal{U}([0, 1])$ si F est continue, et que F continue équivaut à $F(\mathbb{R}) \supset]0, 1[$.

Exercice 14 Comment peut-on simuler une variable de loi uniforme dans un pavé $[0, r_1] \times \dots \times [0, r_d]$ de \mathbb{R}^d ? On poursuit avec le cas $d = 2$. Dans ce cas, comment simuler une variable de loi uniforme dans un rectangle quelconque, de sommets $A = (a_1, a_2)$, $B = (b_1, b_2)$, $C = (c_1, c_2)$, $D = (d_1, d_2)$? Comment simuler une variable de loi uniforme dans un triangle rectangle de sommets $A = (a_1, a_2)$, $B = (b_1, b_2)$, $C = (c_1, c_2)$? et dans un triangle quelconque de sommets $A = (a_1, a_2)$, $B = (b_1, b_2)$, $C = (c_1, c_2)$? Programmez une fonction qui, en fonction des variables d'entrée A, B, C , effectue ce tirage.

À partir de cette méthode et d'un algorithme de découpage d'un polygone en triangles, on peut construire une fonction permettant de tirer des lois uniformes sur un polygone quelconque. Ceci peut être très utile pour accélérer des méthodes par rejet.

1.6 Générateurs pseudo-aléatoires (facultatif)

L'objectif de cette section est de donner quelques éléments de réponse à la question : comment l'ordinateur produit-il des nombres aléatoires ? Jusqu'ici on a vu qu'on pouvait tout ramener à la fonction `random`, qui renvoie des nombres aléatoires uniformes dans $[0,1]$. Mais comment fonctionne-t-elle ?

Première remarque : si on sait simuler une loi uniforme discrète dans $\{0,1,\dots,m-1\}$ pour m assez grand, alors en divisant par m on aura une bonne approximation d'une loi uniforme sur $[0,1]$.

On cherche donc à produire une suite de nombres $(x_n)_{n \in \mathbb{N}}$ appartenant à $\{0,1,\dots,m-1\}$ dont les propriétés sont proches d'une suite aléatoire. Les méthodes dont on va parler sont basées sur des algorithmes déterministes, construits pour que les résultats soient suffisamment chaotiques pour être indistinguables d'une suite de nombres iid. C'est pourquoi on parle de nombre *pseudo-aléatoires*²

La plupart des générateurs sont construits sur une relation de récurrence $x_{n+1} = f(x_n)$. Ils sont donc nécessairement périodiques, de période maximale m . Si cette période est trop petite, le générateur est assez mauvais. Le cas le plus classique consiste à utiliser une *méthode de congruence simple* :

$$x_n = (ax_{n-1} + b) \pmod{m}$$

pour un bon choix de a, b, m et d'une valeur initiale x_0 , appelée graine ou *seed*³.

Exercice 15 Pour $a = 6, b = 0, m = 25$ et $x_0 = 1$, quelle est la période du générateur précédent ? Que dire de la qualité de ces nombres pseudo-aléatoires ?

Il existe un moyen d'assurer que la période maximale est atteinte, grâce à un théorème de Hull et Dobell (que l'on admettra) :

Théorème 1.6.1 (Hull et Dobell) Soient a, b, m tels que

- (i) b et m sont premiers entre eux,
- (ii) $(a - 1)$ est un multiple de chaque nombre premier qui divise m ,
- (iii) si m est multiple de 4 alors $a - 1$ l'est aussi.

2. Il est aussi possible de générer de l'aléatoire par des processus physiques, en mesurant des fluctuations de température ou de tension par exemple. On ne parlera pas de ces méthodes ici.

3. Ainsi, si on lance le même programme sur deux ordinateurs différents avec la même *seed*, on obtiendra la même suite pseudo-aléatoire. Cela peut être pratique pour reproduire des résultats. Si au contraire on veut que la suite soit toujours différente (ou presque), on peut choisir une *seed* différente à chaque fois, par exemple l'heure du système en ms.

Alors pour tout $x_0 \in \{0, \dots, m-1\}$, la suite définie par la récurrence

$$x_n = (ax_{n-1} + b) \pmod{m}$$

a pour période m .

Exercice 16 Produire des triplets (a,b,m) vérifiant les hypothèses avec m arbitrairement grand.

C'est déjà un bon point : on peut produire des suites de période maximale, grâce à une récurrence très simple à calculer. On a donc une certaine équirépartition, mais cela n'empêche pas qu'il y ait des fortes corrélations, par exemple entre x_i et x_{i+1} ...

Exercice 17 On prend $a = 9, b = 3, m = 256$. Testez quelques termes de la suite à l'aide d'une fonction Python. Ces nombres vous semblent-ils satisfaisants comme nombres aléatoires ?

Tracer les 256 points (x_i, x_{i+1}) dans un nuage de points et commenter.

L'exercice précédent montre que de fortes corrélations peuvent subsister. Une manière d'y remédier est d'utiliser une méthode de congruence avec retard $r \in \mathbb{N}^*$:

$$x_n = (ax_{n-r} + b) \pmod{m}$$

et l'on doit désormais choisir r valeurs initiales (ou bien les générer à partir d'une seule graine x_0 et d'une méthode de congruence simple).

Exercice 18 Tracer 256 points (x_i, x_{i+1}) pour la méthode de congruence avec retard avec $m = 256, a = 5, b = 1, r = 6$. On calculera les termes x_0, \dots, x_5 à l'aide d'une congruence simple avec $m = 8, x_0 = 1, a = 5, b = 1$.

Il existe bien d'autres méthodes de génération pseudo-aléatoire (méthode du carré médian, de congruence avec mélanges, de l'inverse, de registre à décalage, ou même basées sur des calculs rapides de décimales de nombres comme π ...). Leur étude est souvent assez délicate et il n'est pas évident de repérer leurs défauts potentiels au premier coup d'œil. Il est judicieux de les soumettre à des tests statistiques pour évaluer leur qualité, comme le test du χ^2 . On aura l'occasion d'effectuer de tels tests en TP.

Chapitre 2

Convergence des variables aléatoires

Le but de ce chapitre est de rappeler les différents modes de convergence de suites $(X_n)_n$ variables aléatoires, leurs propriétés et relations, et de voir comment les illustrer. L'illustration dont on parle ici ne joue pas forcément le rôle d'exemple qui suit un théorème : c'est aussi un outil pour étudier un modèle et établir des conjectures sur son comportement. Mais l'illustration n'est pas non plus une preuve et une fois la conjecture établie, il restera à la démontrer. Dans toute la suite, **on suppose à chaque fois que l'on sait simuler les variables aléatoires $(X_n)_n$ mais pas forcément que l'on sait simuler la limite X .**

2.1 Rappels

Dans cette section, on rappelle sans preuve¹ les définitions et propriétés fondamentales des modes de convergence pour des suites de variables aléatoires. Dans toute cette section, $(X_n)_n$, $(Y_n)_n$, etc. représenteront des suites de variables aléatoires. Suivant les cas, les variables X_n correspondant à différentes valeurs de n vivront sur différents espaces de probabilité $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, ou au contraire sur un même espace $(\Omega, \mathcal{F}, \mathbb{P})$. Toutes les variables considérées sont à valeurs dans un espace \mathbb{R}^d muni d'une norme $\|\cdot\|$ (dont le choix importe peu : toutes ces normes sont équivalentes). On rappelle d'abord les définitions générales :

Définition 2.1.1 *On dit que :*

- une suite $(X_n)_n$ de variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$ converge presque-sûrement vers une variable X , également sur $(\Omega, \mathcal{F}, \mathbb{P})$, si

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1.$$

- une suite $(X_n)_n$ de variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$ converge en norme L^p (où $p \geq 1$) vers une variable X , également sur $(\Omega, \mathcal{F}, \mathbb{P})$, si

$$\lim_{n \rightarrow \infty} \mathbb{E}(\|X_n - X\|^p) = 0.$$

1. Vous trouverez des preuves de ces résultats dans votre livre favori de probabilités, et des contre-exemples à "toutes" les autres implications dans le livre *Counterexamples in probability* de Stoyanov.

- une suite $(X_n)_n$ de variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$ converge en probabilité vers une variable X , également sur $(\Omega, \mathcal{F}, \mathbb{P})$, si pour tout $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\|X_n - X\| > \epsilon) = 0.$$

- une suite de variables aléatoires $(X_n)_n$ sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ converge en loi vers une variable X sur $(\Omega, \mathcal{F}, \mathbb{P})$, si pour toute fonction continue bornée φ sur \mathbb{R}^p ,

$$\lim_{n \rightarrow \infty} \mathbb{E}(\varphi(X_n)) = \mathbb{E}(\varphi(X)).$$

On note $\xrightarrow{p.s.}$, $\xrightarrow{\mathbb{P}}$, $\xrightarrow{L^p}$, $\xrightarrow{\mathcal{L}}$ respectivement ces quatre modes de convergence.

Remarque 2.1.2 La convergence en loi peut donc s'énoncer pour des variables vivant sur des espaces de probabilité différents puisque la convergence concerne la loi des variables et non les variables elles-mêmes.

On rappelle maintenant les relations générales entre ces différents modes de convergence, qui sont résumées par la Figure 2.1.

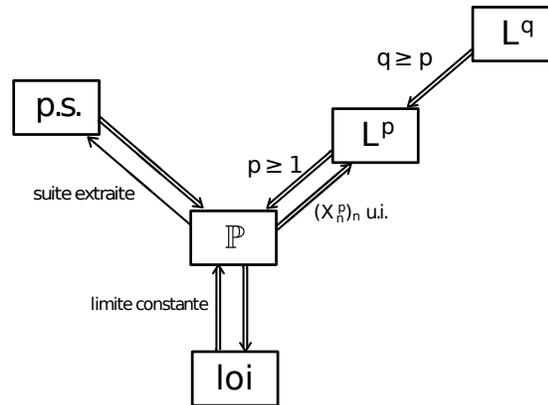


FIGURE 2.1 – La Proposition 2.1.3 en image

Proposition 2.1.3

1. La convergence presque-sûre implique la convergence en probabilité,
2. la convergence L^q implique la convergence L^p si $q \geq p$, et la convergence en probabilité,
3. la convergence en probabilité implique la convergence en loi,
4. la convergence en loi d'une suite $(X_n)_n$ vers une constante c implique la convergence en probabilité vers c
5. la convergence en probabilité implique la convergence presque-sûre d'une suite extraite,

6. la convergence en probabilité d'une suite $(X_n)_n$ vers X est équivalente à la convergence L^1 de $\inf(\|X_n - X\|, 1)$ vers 0,

7. on a équivalence entre les deux points suivants :

- $(X_n)_n$ est uniformément intégrable et $X_n \xrightarrow{\mathbb{P}} X$,
- X est intégrable et $X_n \xrightarrow{L^1} X$.

Rappelons qu'une suite $(X_n)_n$ est uniformément intégrable (ou u.i.) si et seulement si elle vérifie l'une des deux conditions équivalentes suivantes :

$$\limsup_{c \rightarrow \infty} \mathbb{E}(|X_n| \mathbf{1}_{|X_n| > c}) = 0$$

$$\forall \epsilon > 0, \exists \delta > 0 \text{ tel que } \mathbb{P}(A) < \delta \Rightarrow \sup_n \mathbb{E}(|X_n| \mathbf{1}_A) < \epsilon.$$

Il est immédiat que s'il existe Y intégrable telle que pour tout n , on a $|X_n| \leq Y$ p.s. alors $(X_n)_n$ est u.i.

Proposition 2.1.4 *Si f est une fonction continue, alors :*

- Si $X_n \xrightarrow{p.s.} X$, on a $f(X_n) \xrightarrow{p.s.} f(X)$,
- si $X_n \xrightarrow{\mathbb{P}} X$, on a $f(X_n) \xrightarrow{\mathbb{P}} f(X)$,
- Si $X_n \xrightarrow{L^1} X$, on a $f(X_n) \xrightarrow{L^1} f(X)$.

Ces résultats sont utiles si on les combine avec la proposition suivante, puisqu'ils permettront de discuter par exemple la convergence de $(X_n + Y_n)_n$ si l'on suppose la convergence de $(X_n)_n$ et de $(Y_n)_n$:

Lemme 2.1.5

- Si $X_n \xrightarrow{p.s.} X$ et $Y_n \xrightarrow{p.s.} Y$ alors $(X_n, Y_n) \xrightarrow{p.s.} (X, Y)$,
- si $X_n \xrightarrow{\mathbb{P}} X$ et $Y_n \xrightarrow{\mathbb{P}} Y$ alors $(X_n, Y_n) \xrightarrow{\mathbb{P}} (X, Y)$.

Remarque 2.1.6 *Pour la convergence en loi, il n'y a pas de résultat aussi simple, et en particulier $X_n \xrightarrow{L^1} X$ et $Y_n \xrightarrow{L^1} Y$ n'implique pas $X_n + Y_n \xrightarrow{L^1} X + Y$, ni $X_n \times Y_n \xrightarrow{L^1} X \times Y$. La raison en est que la loi de $X + Y$, par exemple, ne dépend pas que des lois de X et de Y mais de la loi du couple (X, Y) . Considérez par exemple $X_n \sim \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_{+1}$ et $Y_n = -X_n$, auquel cas $X_n \xrightarrow{L^1} X_1$, $Y_n \xrightarrow{L^1} X_1$ mais $X_n + Y_n = 0$ ne converge pas en loi vers $2X_1$.*

Un résultat particulièrement utile dans cette direction est le lemme de Slutsky :

Lemme 2.1.7 (lemme de Slutsky) *Si $X_n \xrightarrow{L^1} X$ et $Y_n \xrightarrow{L^1} c$ où c est une constante, alors $(X_n, Y_n) \xrightarrow{L^1} (X, c)$.*

2.2 Illustration de la convergence presque-sûre

La convergence presque-sûre est la plus simple à illustrer. Pour cela, on répète plusieurs fois l'opération suivante :

- on simule quelques réalisations de la suite $(X_n)_n$ pour $n \leq n_{\max}$,
- pour chaque réalisation, on trace la suite des valeurs obtenues.

Si la convergence a lieu, les tracés doivent tous avoir une asymptote. Cependant :

- Il faut faire attention à ce que, pour une réalisation de la suite $(X_n)_n$, chaque $X_n(\omega)$ corresponde au même ω , et que l'on ne change pas l'aléa pour chaque n . L'exercice 19 illustre la différence entre les deux situations.
- On a l'habitude d'illustrer la convergence presque-sûre dans des cas du type "loi des grands nombres" auquel cas la limite est une constante et les trajectoires ont toutes la même asymptote. Il faut faire attention au fait que si $X_n \xrightarrow{\text{p.s.}} X$ mais que la variable limite X n'est pas presque-sûrement constante, la limite peut dépendre du ω , donc l'asymptote n'est pas forcément la même pour toutes les réalisations de suites (voir les exercices 19 et 25). En revanche il n'est pas nécessaire de savoir simuler X pour illustrer $X_n \xrightarrow{\text{p.s.}} X$, et l'on peut même estimer la loi de X grâce à ces simulations.

Ce que l'on a écrit ci-dessus pose plusieurs questions :

combien de tracés ? On veut illustrer le fait que $p := \mathbb{P}(\omega \mid X_n(\omega) \text{ converge})$ est égale à 1. Si l'on fait M tracés indépendants, la probabilité de n'avoir "choisi" que des ω pour lesquels on a convergence est p^M . On peut formaliser la situation par un test statistique, dans lequel on souhaite distinguer entre $H_0 : p = 1$ et $H_1 : p < 1$, et on rejette H_0 si on observe une réalisation sans convergence parmi les M réalisations indépendantes. On peut considérer que l'erreur de première espèce est 0, car on ne rejettera jamais H_0 à tort (pourvu que les trajectoires soient assez longues pour constater la convergence, voir le point suivant), et la puissance du test est $1 - p^M$. On choisira alors M en fonction de la puissance souhaitée.

Notons que dans de nombreuses situations (en particulier quand une loi du 0-1 s'applique) on pourra prouver que la probabilité de convergence est soit 0, soit 1 ; dans ce cas il suffira d'observer la convergence dans une situation pour avoir une indication de la convergence presque-sûre.

quelle longueur de trajectoire ? autrement dit, jusqu'à quelles valeurs de n_{\max} pousser ? Il n'y a pas de bonne réponse puisque cela dépend de la vitesse de convergence, que l'on ne connaît pas (sauf dans le cas où l'on cherche simplement à illustrer un résultat). L'erreur que l'on risque de commettre par un mauvais choix de n_{\max} est de conclure à l'absence de convergence, alors que la convergence est seulement trop lente.

peut-on en tirer une estimation de la vitesse de convergence ? on pourra tenter d'intuiter la vitesse de convergence en utilisant des tracés modifiés : si un tracé log-log (c'est-à-dire un tracé de $\log(X_n - X)$ en fonction de $\log n$) a une apparence affine, alors la convergence semble être polynomiale

(d'exposant égal à la pente négative), si un tracé logarithmique a une apparence affine, la convergence semble être exponentielle (de taux égal à la pente négative)...

Exercice 19 Soit $(R_n)_{n \in \mathbb{N}}$ une suite i.i.d. de variables de Rademacher $\mathbb{P}(R_0 = \pm 1) = 1/2$. Pour un $a \in]-1, +1[$ fixé on pose

$$X_n = a^0 R_0 + \dots + a^{n-1} R_{n-1}.$$

Montrez que la suite X_n converge presque-sûrement. Supposons maintenant que pour la simulation de chaque X_n , on tire à nouveau les R_0, \dots, R_{n-1} , de sorte que ce qu'on calcule est

$$X'_n = a^0 R'_{n,0} + \dots + a^{n-1} R'_{n,n-1}$$

où les $(R'_{i,j})$ sont iid de même loi. A-t-on à nouveau convergence presque-sûre ? et pour

$$X''_n = a^0 R_{n-1} + \dots + a^{n-1} R_0 ?$$

2.3 Illustration de la convergence en loi

Il existe plusieurs méthodes pour illustrer la convergence en loi, dont les plus classiques sont :

- dans le cas de variables aléatoires discrètes, le tracé de diagrammes en bâtons,
- dans le cas de variables à densité, le tracé des densités,
- le tracé d'histogrammes,
- le tracé de fonctions de répartitions empiriques.

La première méthode est basée sur le résultat suivant :

Proposition 2.3.1 Si $(X_n)_n$ est une suite de variables aléatoires à valeurs dans un ensemble discret D , alors $(X_n)_n$ converge en loi si et seulement si pour tout $d \in D$, la limite $\mu_d = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = d)$ existe et que $\sum_d \mu_d = 1$. Dans ce cas, la loi limite est portée par D et décrite par les μ_d .

Dans la suite, nous parlerons principalement de la méthode par les fonctions de répartition empiriques, qui a deux avantages notables : elle fonctionne dans tous les cas où l'on travaille avec des variables réelles, et elle est simple à coder.

Proposition 2.3.2 Soit $(X_n)_n$ une suite de variables aléatoires réelles. Alors on a $X_n \xrightarrow{L} X$ si et seulement si $F_{X_n}(t) \rightarrow F_X(t)$ pour tout t point de continuité de F_X .

Utiliser le résultat ci-dessus suppose de connaître la loi de X . Si tout ce que l'on observe est une convergence ponctuelle de $(F_{X_n})_n$ vers une fonction F , alors on n'a pas forcément convergence en loi : il faut en plus que F tende vers 0 en $-\infty$

et vers 1 en $+\infty$, ce qui n'est pas forcément facile à identifier. Dans ce cas, F sera la fonction de répartition d'une unique loi.

Remarquons que même si l'on arrive à simuler X_n , on n'a pas forcément immédiatement accès à F_{X_n} . Pour approximer cette fonction, on utilisera le théorème de Glivenko-Cantelli :

Théorème 2.3.3 (Glivenko-Cantelli) *Soit Y une variable aléatoire réelle et $(Y^{(k)})_{k=1}^N$ est un N -échantillon de même loi que Y . Soit $\widehat{F}_Y^{(N)}$ la fonction de répartition empirique :*

$$\widehat{F}_Y^{(N)}(t) = \frac{1}{N} \sum_{k=1}^N \mathbb{1}_{Y^{(k)} \leq t}.$$

Alors $\widehat{F}_Y^{(N)}$ converge presque-sûrement uniformément en t vers la fonction de répartition F_Y de Y . Autrement dit,

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} |\widehat{F}_Y^{(N)}(t) - F_Y(t)| \xrightarrow{N \rightarrow \infty} 0\right) = 1.$$

Ce théorème, s'il est un peu pénible à démontrer, est essentiellement une application de la loi des grands nombres. Pour avoir une bonne estimation de F_Y , encore faut-il choisir N assez grand. Une estimation peut être donnée par le théorème de Kolmogorov-Smirnov :

Théorème 2.3.4 (Kolmogorov-Smirnov) *Si de plus F_Y est continue,*

$$\sqrt{N} \sup_{t \in \mathbb{R}} |\widehat{F}_Y^{(N)}(t) - F_Y(t)|$$

converge en loi quand $N \rightarrow \infty$, vers une loi appelée loi de Kolmogorov-Smirnov².

On reconnaît un théorème universel, comme le théorème central limite : la loi limite est la même quel que soit Y . Comme la loi de Kolmogorov-Smirnov est intégrable (elle a même des moments de tous ordres), le sup en question est de l'ordre de $1/\sqrt{N}$. Ainsi, si on veut approcher F_Y uniformément sur un segment avec précision ϵ , on prendra N d'ordre $1/\epsilon^2$.

Par ailleurs, le théorème de Kolmogorov-Smirnov fournit des tests efficace pour savoir si un échantillon suit bien une loi Y donnée, ou encore si deux échantillons suivent la même loi (de loi possiblement inconnue).

Exercice 20 *Soit $(X_n)_n$ une suite de variables i.i.d. suivant une loi qui admet deux premiers moments, et que vous savez simuler. Illustrez le théorème de la limite centrale.*

Exercice 21 *Dans l'exercice 19 montrez que les trois suites $(X_n)_n$, $(X'_n)_n$, $(X''_n)_n$ convergent en loi et illustrez cette convergence.*

2. Sa fonction de distribution est $\mathbb{P}(K \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}$.

L'exercice suivant commence l'étude d'un classique appelé le “*collectionneur de vignettes*”³.

Exercice 22 Soit $(Y_k)_k$ une famille i.i.d. de variables de loi uniforme sur $\{1, \dots, n\}$. On note T_n le premier instant où les Y ont pris toutes les valeurs possibles :

$$T_n = \inf \{k \text{ tel que } \#\{Y_1, \dots, Y_k\} = n\}.$$

Écrire une fonction Python simulant T_n ; on pourra utiliser la structure de données appelée `set` (qui représente donc un ensemble). Expérimentez pour voir si $T_n / (n^\alpha \log^\beta n)$ semble converger en loi, pour des valeurs simples de α, β .

2.4 Illustration de la convergence \mathbb{P}

Compte tenu des points 3 et 4 de la Proposition 2.1.3, le plus simple pour illustrer la convergence en probabilité de $(X_n)_n$ vers X est d'illustrer la convergence en loi de $X_n - X$ vers zéro. On peut également chercher à montrer la convergence L^1 vers zéro de $\inf(|X_n - X|, 1)$ avec les méthodes développées dans la section suivante. Dans les deux cas il faut en général connaître X pour appliquer ces méthodes.

2.5 Illustration des convergences L^p

La méthode naturelle pour illustrer le fait $\mathbb{E}(|X_n - X|^p) \rightarrow 0$ est d'utiliser les moyennes empiriques :

- on simule, pour différents n et N grand, N réalisations $(X_n^{(k)} - X^{(k)})_{k=1}^N$ de $X_n - X$,
- on calcule la moyenne empirique $\frac{1}{N} \sum_{k=1}^N |X_n^{(k)} - X^{(k)}|^p$,
- on trace la suite en n de ces quantités.

Si la convergence a lieu, on s'attend à avoir convergence vers zéro de cette suite. Encore une fois, plusieurs questions se posent :

pour quels n ? on ne peut donner que la même réponse qu'en section 2.2 : ça dépend, et on fait comme on peut.

pour quelle(s) taille(s) N d'échantillon ? ce choix ne peut se faire que compte tenu d'un seuil de confiance choisi par ailleurs.

On a deux problèmes cependant :

- On veut ici des intervalles de confiance pour des quantités dont on cherche à montrer qu'elles tendent vers zéro : il va donc falloir pouvoir faire tendre le diamètre de l'intervalle de confiance vers zéro aussi, et pour que cela soit réalisable en pratique, il faudra que le N ne croisse pas trop vite.
- En général, le N permettant une précision donnée dans l'estimation de $\mathbb{E}(|X_n - X|^p)$ va dépendre de n . Puisque l'on veut illustrer la convergence des quantités estimées, on voudrait idéalement pouvoir choisir N tel que les intervalles de confiance donnés sont valables pour tout n “assez grand”.

3. le terme anglais étant “coupon collector”, on a tendance à parler de “collectionneur de coupons”

Il n'y a pas de bonne réponse à ces problèmes ; nous verrons quelques outils plus tard dans le cours. En pratique on pourra souvent illustrer une autre forme de convergence plutôt qu'une convergence L^p : si par exemple on suppose les $(X_n)_n$ à valeurs dans $[a, b]$ alors l'inégalité de Hoeffding permet de répondre de manière satisfaisante aux deux points ci-dessous, mais alors $(X_n)_n$ est uniformément intégrable et, donc $X_n \xrightarrow{L^1} X$ équivaut à $X_n \xrightarrow{p.s.} X$ ou à $X_n \xrightarrow{\mathbb{P}} X$, qui équivaut à $X_n - X \xrightarrow{L^1} 0$. Remarquons par ailleurs qu'il est nécessaire de savoir simuler X (ou en tout cas $X_n - X$) pour appliquer ces méthodes.

Exercice 23 Soit $(S_n)_n$ une marche aléatoire symétrique sur \mathbb{Z} et $X_n = \mathbf{1}_{S_n=0}$. Montrer que $X_n \xrightarrow{L^1} 0$, et rappeler pourquoi que $X_n \not\xrightarrow{p.s.} 0$. Essayez d'illustrer la convergence L^1 par la méthode décrite ci-dessus (avec un N fixé a priori pour toutes les valeurs de n entre 1 et n_{\max}).

2.6 Exercices

Exercice 24 Dans les cas suivants, démontrer et illustrer les convergences ou absences de convergence proposées.

1. Si $X_n \sim \mathcal{B}(n, \frac{\lambda}{n})$ avec $\lambda \in \mathbb{R}_+$, alors $X_n \xrightarrow{L^1} \mathcal{P}(\lambda)$.
2. Si les $(X_n)_{n \in \mathbb{N}}$ sont i.i.d. avec $X_0 \sim \mathcal{B}(\frac{1}{2})$, et si $Y_n = \sum_{k=0}^n X_k 2^{-k}$, alors $Y_n \xrightarrow{p.s.} \mathcal{U}([0, 1])$.
Que dire si $X_0 \sim \mathcal{B}(p)$ où $p \in [0, 1]$?
3. Si $X_n \sim \mathcal{B}(\frac{1}{n})$, alors $X_n \xrightarrow{\mathbb{P}} 0$, $X_n \xrightarrow{L^1} 0$, mais $X_n \not\xrightarrow{p.s.} 0$. Et pour nX_n ?

Exercice 25 (inspiré du texte d'agrégation public 2015-A7) On définit deux variables aléatoires A_n par $A_0 = a$, $B_0 = b$ et

$$\mathbb{P}((A_{n+1}, B_{n+1}) = (A_n + 1, B_n) \mid (A_n, B_n)) = \frac{A_n}{n + a + b}$$

$$\mathbb{P}((A_{n+1}, B_{n+1}) = (A_n, B_n + 1) \mid (A_n, B_n)) = \frac{B_n}{n + a + b},$$

(le processus $(A_n, B_n)_n$ est donc une urne de Pólya).

- Simuler une réalisation de la suite $(A_n)_n$. Semble-t-il y avoir convergence presque-sûre de A_n/n ?
- On se place dans le cas $a = b = 1$. Quelle loi semble suivre la limite de A_n/n ? On pourra tracer une approximation de la fonction de répartition de la loi de A_n/n à partir d'un échantillon, pour n assez grand.
- On conjecture que la loi limite de A_n/n est une loi $\beta(a, b)$; illustrer ce point.

On peut prouver la convergence presque-sûre en montrant que $\frac{A_n + a}{n + a + b}$ est une martingale.

Exercice 26 On reprend l'exercice 22 mais cette fois-ci le collectionneur ne cherche qu'à obtenir une proportion $\rho \in]0,1[$ des vignettes disponibles, et on note $T_n^{(\rho)}$ le temps correspondant. Montrer (par la simulation) que pour tout $\rho \in]0,1[$ il existe deux constantes m_ρ et σ_ρ^2 pour lesquelles $\frac{T_n^{(\rho)} - m_\rho n}{\sqrt{\sigma_\rho^2 n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

Pour cela, on commencera par observer que $\mathbb{E}(T_n)$ et $\text{var}(T_n)$ sont approximativement linéaires en n . On pourra alors estimer les coefficients m_ρ et σ_ρ^2 à partir de moyenne et variance empiriques de $T_n^{(\rho)}$ pour n assez grand.

Chapitre 3

Grands théorèmes de convergence

Le but de ce chapitre est de rappeler les grands théorèmes de convergence de variables aléatoires, et de donner des méthodes permettant de traiter au cas par cas les situations où les hypothèses de ces théorèmes ne sont pas vérifiées. Les deux types de résultats de convergence que nous discuterons sont les suivants (à chaque fois $(a_n)_n$ et $(b_n)_n$ sont des suites déterministes) :

- les lois des grands nombres ou théorèmes ergodiques, qui sont du type “ $\frac{1}{a_n} \sum_{k=1}^n X_k$ converge vers une quantité déterministe”,
- les résultats du type “central limite” qui disent que “ $\frac{1}{b_n} \sum_{k=1}^n (X_k - \mathbb{E}(X_k))$ converge en loi”,

Tous les livres classiques de probabilités discutent loi des grands nombres et théorème central limite, mais tous ne discutent pas leurs variantes. Une lecture très intéressante sur ces points se trouve sur le blog de Terence Tao <https://terrytao.wordpress.com> (chercher “275A probability theory” ; il y a six chapitres au total, “Notes 0” à “Notes 5”).

3.1 Lois des grands nombres

Dans toute la suite, on considère une suite de variables aléatoires $(X_n)_{n \in \mathbb{N}^*}$. On notera

$$S_n = X_1 + \dots + X_n$$

Un résultat de loi des grands nombres concerne la convergence de S_n/a_n vers une constante (en général $a_n = n$ mais pas toujours, voir l'exercice 31). On parlera de loi forte pour un résultat de convergence presque-sûre, et de loi faible pour un résultat de convergence en probabilité.

Le résultat le plus classique est la *loi forte des grands nombres* :

Théorème 3.1.1 (Kolmogorov) *Soit $(X_n)_n$ une suite i.i.d. de variables aléatoires. Alors la suite $(\bar{X}_n)_n$ converge p.s. si et seulement si $X_1 \in L^1$, et alors la limite est $\mathbb{E}(X_1)$.*

La démonstration n'est pas évidente, mais il faut savoir traiter des cas où les hypothèses sont plus fortes :

Exercice 27 On cherche à démontrer un sens du Théorème 3.1.1 dans le cas où $X_1 \in L^4$ (c'est-à-dire $\mathbb{E}[X_1^4] < \infty$). On pourra supposer X_1 centrée (pourquoi ?).

1. Calculer $\mathbb{E}[S_n^4]$ en fonction de $\sigma^2 = \mathbb{E}[X_1^2]$ et $\tau^4 = \mathbb{E}[X_1^4]$.
2. Montrer que $\mathbb{E}[\bar{X}_n^4] \leq \frac{3\tau^4}{n^2}$.
3. En déduire que $\mathbb{E}[\sum_{n=1}^{\infty} \bar{X}_n^4] < \infty$.
4. Conclure la convergence p.s. de \bar{X}_n vers 0.

On affaiblit les hypothèses en supposant seulement $X_1 \in L^2$. Montrer la convergence en probabilité, en utilisant l'inégalité de Bienaymé-Tchebychev.

Il existe d'autres résultats du type "loi des grands nombres" dans les cadres suivants :

- pour les chaînes de Markov : c'est le théorème ergodique, qui donne une convergence presque-sûre de toute quantité $\frac{1}{n} \sum_{k=1}^n f(X_k)$ dès que la chaîne admet une unique mesure invariante π , dès que f est une fonction π -intégrable (la limite étant $\int f d\pi$);
- pour les martingales : si $(M_n)_n$ une martingale telle que chaque M_n est de carré intégrable, alors il existe un processus croissant $\langle M \rangle_n$ tel que sur l'événement $[\lim_n \langle M \rangle_n = \infty]$, on a $\frac{M_n}{\langle M \rangle_n} \xrightarrow{\text{p.s.}} 0$.

Revenons au Théorème 3.1.1. Il y a trois hypothèses que l'on aimerait affaiblir :

1. les $(X_n)_n$ ont même loi,
2. les $(X_n)_n$ sont indépendantes,
3. les $(X_n)_n$ sont intégrables.

Le Théorème 3.1.1 étant un "si et seulement si", aucun espoir d'avoir une convergence presque-sûre de S_n/n sans l'**hypothèse 3** : l'exercice suivant illustre cela.

Exercice 28 Comment se comporte (au sens presque-sûr) S_n/n quand les $(X_n)_n$ suivent des lois de Cauchy ? et au sens de la convergence en loi ? Conjecturer le résultat par simulation, puis le prouver (indication : la fonction caractéristique de X_1 est $t \mapsto e^{-|t|}$).

Notons qu'il existe un analogue faible qui permet d'affaiblir l'**hypothèse 3** :

Théorème 3.1.2 Soit $(X_n)_n$ une suite i.i.d. de variables aléatoires. Alors la suite $(\frac{1}{n}(S_n - n\mathbb{E}(X_1 \mathbf{1}_{|X_1| \leq n})))_n$ converge en probabilité si et seulement si $\lim_{t \rightarrow \infty} \mathbb{P}(|X_1| \geq t) = 0$.

Pour ce qui est d'affaiblir l'**hypothèse 1**, le théorème des trois séries de Kolmogorov (encore lui) montre que si les $(X_n)_n$ sont indépendantes et toutes de carré intégrable,

$$\text{si } \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k) \rightarrow \mu \quad \text{et} \quad \sum_n \frac{\text{var } X_n}{n^2} < \infty \quad \text{alors} \quad \bar{X}_n \xrightarrow{\text{p.s.}} \mu.$$

Ce résultat est une conséquence des lois des grands nombres pour les martingales.

Un premier résultat affaiblissant l'hypothèse 2 est simple (on pourra le trouver dans les livres de Barbe et Ledoux ou de Bercu et Chafaï)

Théorème 3.1.3 (Etemadi) *Soit $(X_n)_n$ une suite de variables aléatoires deux à deux indépendantes. Alors la suite $(\bar{X}_n)_n$ converge presque-sûrement si et seulement si $X_1 \in L^1$, et alors la limite est $\mathbb{E}(X_1)$.*

Il existe aussi des résultats pour les variables « faiblement corrélées », comme ce théorème de Lyons qu'on donne aussi pour la culture :

Théorème 3.1.4 (Lyons) *Soit $(X_n)_n$ une suite de variables aléatoires bornées et telles que*

$$\text{Cov}(X_n, X_m) \leq \phi(|n - m|)$$

où ϕ est une fonction telle que

$$\sum_{n \geq 1} \frac{\phi(n)}{n} < \infty.$$

Alors la loi forte des grands nombres s'applique à (X_n) .

Des méthodes “à la main” permettant d'affaiblir l'hypothèse 2 peuvent aussi être obtenues — quitte à renforcer un peu les hypothèses — en utilisant la **méthode des moments**, comme on l'a fait dans l'exercice 27. En voici quelques exemples.

Exercice 29 *En s'inspirant de l'exercice 27, montrer le résultat suivant :*

Si $(X_{i,n})_{i \leq n}$ est une famille de variables aléatoires qui ont toutes la même espérance $\mathbb{E}(X_{1,1})$ et telle que pour tout n , $X_{1,n}, \dots, X_{n,n}$ est une famille indépendante et $\sup_{i,n} \mathbb{E}(X_{i,n}^p) < \infty$, alors $\frac{1}{n} \sum_{k=1}^n X_{k,n}$ converge vers $\mathbb{E}(X_{1,1})$ en probabilité si $p = 2$ et presque-sûrement si $p = 4$.

Exercice 30 *On fabrique une suite $(V_n, E_n)_{n \in \mathbb{N}^*}$ de graphes aléatoires dits de Erdős–Rényi de la manière suivante :*

- l'ensemble des sommets V_n est $\{1, \dots, n\}$,
- chaque $\{a, b\}$ (où $a \neq b$) est une arête du graphe avec probabilité $1/2$, et indépendamment des autres arêtes.

Écrire une fonction Python simulant un tel graphe ; on pourra pour cela coder le graphe par sa matrice d'adjacence.

Montrer que $\#E_n / \binom{n}{2}$ converge presque-sûrement vers $1/2$.

Exercice 31 *Soit $(Y_k)_k$ une famille i.i.d. de variables de loi uniforme sur $\{1, \dots, n\}$. On note T_n le premier instant où les Y ont pris toutes les valeurs possibles :*

$$T_n = \inf \{k \text{ tel que } \#\{Y_1, \dots, Y_k\} = k\}.$$

On rappelle qu'on a conjecturé la convergence en probabilité de $T_n/n \log n$ vers 1 ; montrer que $T_n = X_{n,1} + \dots + X_{n,n}$ pour des variables aléatoires indépendantes de loi géométrique, calculer l'espérance de T_n et majorer sa variance. Montrer alors que $T_n/n \log n \xrightarrow{\mathbb{P}} 1$.

Exercice 32 (*) On choisit une permutation σ_n aléatoirement et de manière uniforme dans \mathfrak{S}_n et on s'intéresse au nombre C_n de cycles dans cette permutation. On codera une permutation de la manière suivante :

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 9 & 6 & 8 & 2 & 1 & 5 & 4 & 7 \end{pmatrix}$$

est représenté par $\text{sigma}=[3,9,6,8,2,1,5,4,7]$.

Écrire une fonction qui prend en entrée une permutation et rend une décomposition en cycles.

Observez empiriquement que l'espérance et la variance de C_n sont de l'ordre de $\log n$.

On peut prouver en fait que C_n a même loi que la somme de $X_{n,k}$ indépendants, pour $k = 1, \dots, n$ où $X_{n,k} \sim \mathcal{B}(\frac{1}{n-k+1})$; voir Probability theory and examples de Durrett. En acceptant ce résultat, prouvez les résultats observés.

Montrez ensuite que $C_n/\log n$ tend vers 1 en probabilité. L'illustrer est assez difficile à cause de la lenteur de la convergence.

3.2 Théorèmes centraux limite

Ici aussi, nous allons discuter l'énoncé standard et les extensions que l'on peut obtenir en améliorant ses différentes méthodes de preuve. Nous allons aussi discuter la vitesse de convergence.

Théorème 3.2.1 Si $(X_n)_n$ est une suite i.i.d. de variables réelles de carré intégrable, alors

$$\frac{S_n - n\mathbb{E}(X_1)}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{var}(X_1)).$$

La preuve la plus courante et la plus courte utilise **les fonctions caractéristiques**

$$\begin{aligned} \Phi_Z : \mathbb{R} &\rightarrow \mathbb{C} \\ t &\mapsto \mathbb{E}(e^{itZ}) \end{aligned}$$

et le théorème de Lévy-Cramér. Cette preuve est la plus courte mais elle n'est pas celle qui s'étend le plus facilement. Une autre méthode de preuve du théorème de la limite centrale passe par **les moments**. Elle utilise le résultat suivant : si une suite de variables aléatoires $(Y_n)_n$ vérifie $\mathbb{E}(Y_n^p) \rightarrow \mathbb{E}(Z^p)$ pour tout $p \in \mathbb{N}$, où $Z \sim \mathcal{N}(0,1)$ alors $Y_n \xrightarrow{\mathcal{L}} Z$ (et on peut remplacer ici $\mathcal{N}(0,1)$ par toute loi "déterminée par ses moments", i.e. qui est la seule avec les moments donnés — et c'est le cas de la loi normale).

Ce théorème admet une version *multidimensionnelle* :

Théorème 3.2.2 Si $(X_n)_n$ est une suite i.i.d. de variables dans \mathbb{R}^d de carré intégrable, on note $\Sigma \in \mathcal{M}_{d,d}(\mathbb{R})$ leur matrice de covariance :

$$\Sigma_{i,j} = \text{Cov}(X_1^{(i)}, X_1^{(j)}).$$

Alors

$$\frac{S_n - n\mathbb{E}(X_1)}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

c'est-à-dire que la suite de vecteurs aléatoires converge en loi vers un vecteur gaussien¹ centré de matrice de covariance Σ .

Pour la culture, on donne une généralisation du Théorème Central Limite 3.2.1, qui consiste à sommer des variables différentes pour chaque n , comme on l'a fait dans les exercices 29,30,31.

Théorème 3.2.3 (Lindeberg-Feller) Soit K_n une suite d'entiers avec $K_n \rightarrow \infty$ quand $n \rightarrow \infty$, et soit pour tout n une famille $\tilde{X}_{n,1}, \dots, \tilde{X}_{n,K_n}$ de variables aléatoires indépendantes. On note $\tilde{\sigma}_n^2 = \sum_{k=1}^{K_n} \text{var}(\tilde{X}_{n,k})$; alors la condition

$$\forall \epsilon > 0 \quad \frac{1}{\tilde{\sigma}_n^2} \sum_{k=1}^{K_n} \mathbb{E} \left[(\tilde{X}_{n,k} - \mathbb{E}(\tilde{X}_{n,k}))^2 \mathbb{1}_{|\tilde{X}_{n,k} - \mathbb{E}(\tilde{X}_{n,k})| > \epsilon \tilde{\sigma}_n} \right] \xrightarrow{n \rightarrow \infty} 0 \quad (3.1)$$

est vérifiée si et seulement si on a

$$\frac{1}{\tilde{\sigma}_n^2} \max_{1 \leq k \leq K_n} \text{var}(\tilde{X}_{n,k}) \xrightarrow{n \rightarrow \infty} 0$$

et

$$\frac{1}{\tilde{\sigma}_n} \sum_{k=1}^{K_n} (\tilde{X}_{n,k} - \mathbb{E}(\tilde{X}_{n,k})) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Remarque 3.2.4

- On retrouve le résultat standard donné par le Théorème 3.2.1 en prenant $\tilde{X}_{n,k} = X_k$.
- La condition (3.1) est appelée "condition de Lindeberg". Elle signifie qu'aucun $X_{n,k}$ ne joue un rôle prédominant dans la somme $\sum_{k=1}^{K_n} (\tilde{X}_{n,k} - \mathbb{E}(\tilde{X}_{n,k}))$.

Exercice 33 On utilise les notations de l'exercice 32. On pose $\tilde{X}_{n,k} = X_{n,n-k}$; en utilisant le Théorème 3.2.3, montrer que $\frac{C_n - \log n}{\sqrt{\log n}}$ converge en loi vers une variable normale centrée réduite puis illustrer ce résultat.

1. Pour des rappels sur les vecteurs gaussiens, on pourra consulter par exemple *Probabilité* de Barbé et Ledoux, chapitre IV.4.

Remarque 3.2.5 Une condition suffisante pour (3.1) est qu'il existe $\delta > 0$ tel que

$$\lim_{n \rightarrow \infty} \frac{1}{\bar{\sigma}_n^{2+\delta}} \sum_{k=1}^{K_n} \mathbb{E}(|\tilde{X}_{n,k} - \mathbb{E}(\tilde{X}_{n,k})|^{2+\delta}) = 0. \quad (3.2)$$

Cette condition (3.2) est appelée "condition de Lyapounov".

Exercice 34 On reprend l'exercice 31 mais cette fois-ci le collectionneur ne cherche qu'à obtenir une proportion $\rho \in]0,1[$ des vignettes disponibles.

Montrer que le temps $T_n^{(\rho)}$ peut s'écrire comme $X_{n,1} + \dots + X_{n,r_n}$ avec $r_n \sim \rho n$ et les $(X_{n,k})_{k=1}^{r_n}$ indépendants avec $X_{n,k}$ géométrique de paramètre $1 - k/n$.

Utiliser le Théorème 3.2.3 pour montrer que pour tout $\rho \in]0,1[$ il existe deux constantes m_ρ et σ_ρ^2 pour lesquelles $\frac{T_n^{(\rho)} - m_\rho n}{\sqrt{\sigma_\rho^2 n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. On pourra prouver que la condition (3.2) est vérifiée avec $\delta = 2$ en utilisant l'inégalité²

$$\mathbb{E} \left(\left| G_p - \frac{1}{p} \right|^4 \right) \leq \frac{K}{p^4}$$

pour G_p suivant une loi géométrique de paramètre p , et K une certaine constante (indépendante de p).

Et pour $\rho = 1$? On rappelle que $E(T_n) \sim n \log n$ et que $\text{var}(T_n) \sim \frac{\pi^2}{6} n^2$; a-t-on convergence en loi de $\frac{T_n - n \log n}{\pi n / \sqrt{6}}$?

Parlons maintenant de vitesse de convergence dans le théorème central limite. Le résultat principal est le théorème suivant :

Théorème 3.2.6 (Berry-Esséen) Soit $(X_n)_n$ une suite de variables aléatoires i.i.d. admettant des moments d'ordre 3. On note F_n la fonction de répartition de $Z_n = \frac{S_n - n\mathbb{E}(X_1)}{\sqrt{n \text{var}(X_1)}}$ et F la fonction de répartition de $\mathcal{N}(0,1)$. Alors

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{C\rho}{\sigma^3 \sqrt{n}}$$

où $C \simeq 0,4748$ est une constante universelle et $\rho = \mathbb{E}(|X_1 - \mathbb{E}(X_1)|^3)$.

Ce résultat exprimé en termes de fonctions de répartition donne en fait une estimation de $\mathbb{E}(\phi(Z_n)) - \mathbb{E}(\phi(Z))$ (où $Z \sim \mathcal{N}(0,1)$) pour toute fonction suffisamment régulière, via les identités

$$\mathbb{E}(\phi(Z_n)) = \int \phi'(t)(1 - F_n(t)) dt \quad \mathbb{E}(\phi(Z)) = \int \phi'(t)(1 - F(t)) dt$$

Ce résultat montre donc essentiellement que la vitesse de convergence dans le théorème central limite, Théorème 3.1.1, est en $O(1/\sqrt{n})$.

2. facile à montrer si l'on connaît les moments d'ordre 3 et 4 de G_p .

Exercice 35 Soit $(X)_n$ une suite de variables i.i.d. de Rademacher ; l'inégalité du Théorème 3.2.6 est censée être atteinte pour ces variables. On note $F_n^{(N)}$ la fonction de répartition empirique approchant F_n associée au N -échantillon.

1. Montrer que $\sup_{x \in \mathbb{R}} |F_n^{(N)}(x) - F(x)|$ est atteint en l'un des points de l'échantillon.
2. Donner une estimation de cette quantité pour différentes valeurs de n et la comparer au majorant donné dans le Théorème 3.2.6.

Un autre raffinement du Théorème Central Limite consiste à étudier les écarts records de $\frac{S_n - n\mathbb{E}(X_1)}{\sqrt{n}}$. Le théorème suivant montre moralement qu'ils prennent des valeurs de l'ordre de $\log \log n$ une infinité de fois.

Théorème 3.2.7 (Loi du log itéré) Soit $(X_n)_n$ une suite i.i.d. de carré intégrable, et

$$V_n = \frac{1}{\sqrt{2n \log \log n}} (S_n - n\mathbb{E}(X_1)).$$

Alors presque-sûrement

$$\limsup_{n \rightarrow \infty} V_n = +\sqrt{\text{var}(X_1)} \quad \liminf_{n \rightarrow \infty} V_n = -\sqrt{\text{var}(X_1)}.$$

Exercice 36 A-t-on convergence presque-sûre ou même en proba dans le théorème de la limite centrale ? Pour justifier théoriquement la réponse, on utilisera le Théorème 3.2.7. Pourquoi ne peut-on pas en pratique illustrer ce résultat ?

On peut en revanche observer que V_n prend des valeurs "éloignées de zéro" pour des temps arbitrairement grands. Simuler une trajectoire de V_n pour des X_n de loi de Rademacher.

3.3 Valeurs extrêmes

Soit encore $(X_n)_n$ une famille i.i.d. de variables réelles. Nous nous intéressons ici à la variable donnant les maxima de la suite :

$$M_n = \max_{k=1, \dots, n} X_k$$

et à l'éventuelle convergence en loi de $(M_n - a_n)/b_n$ quand $n \rightarrow \infty$.

On peut commencer par se demander quel est le comportement de M_n lui-même.

Exercice 37 Soit $x_F = \sup\{x \in \mathbb{R} \mid \mathbb{P}(X_1 \leq x) < 1\} \in]-\infty, +\infty]$. Montrer que $M_n \xrightarrow{\mathbb{P}} x_F$.

La question de la convergence de $(M_n - a_n)/b_n$ consiste alors à étudier le comportement de $x_F - M_n$ (si $x_F < \infty$) ou la vitesse de divergence de M_n (si $x_F = \infty$). Le résultat standard en la matière est le suivant :

Théorème 3.3.1 (Fisher-Fréchet-Gnedenko-Tippett) Soit $(X_n)_n$ une famille i.i.d. de variables réelles. S'il existe deux suites, $(a_n)_n$ de réels quelconques et $(b_n)_n$ de réels strictement positifs, telles que $(M_n - a_n)/b_n$ converge en loi quand $n \rightarrow \infty$, alors cette loi limite est nécessairement, à translation et dilatation près, de l'un des quatre types suivants :

- une masse de Dirac,
- une loi de Gumbel, i.e. de fonction de répartition $x \mapsto e^{-e^{-x}}$,
- une loi de Weibull i.e. de fonction de répartition $x \mapsto e^{-(x)^a} \mathbb{1}_{\mathbb{R}_+}(x) + \mathbb{1}_{\mathbb{R}_-}(x)$ avec $a > 0$,
- une loi de Fréchet, i.e. de fonction de répartition $x \mapsto e^{-x^{-a}} \mathbb{1}_{\mathbb{R}_+}(x)$ avec $a > 0$.

Nous n'allons pas démontrer ce résultat mais l'illustrer dans des cas correspondant à différentes situations.

Exercice 38 Dans les cas suivants, prouvez puis illustrez le résultat.

1. Si $X_1 \sim \mathcal{B}(p)$ avec $p \in]0,1[$, on a $M_n \xrightarrow{\mathcal{L}} \delta_1$.
2. Si $X_1 \sim \mathcal{U}([0,\theta])$ pour $\theta > 0$, on a $(M_n - \theta)/\frac{\theta}{n} \xrightarrow{\mathcal{L}}$ une loi de Weibull avec $a = 1$ (qui est la même chose que la loi de $-E$ pour $E \sim \mathcal{E}(1)$).
3. Si $X_1 \sim \mathcal{E}(\lambda)$ avec $\lambda > 0$, on a $(M_n - \frac{\log n}{\lambda})/\frac{1}{\lambda} \xrightarrow{\mathcal{L}}$ une loi de Gumbel.
4. Si $X_1 \sim \mathcal{C}(1)$ alors $M_n/\frac{n}{\pi} \xrightarrow{\mathcal{L}}$ une loi de Fréchet avec $a = 1$.

3.4 Principes de grandes déviations

Pour une suite (X_n) de variables aléatoires i.i.d. réelles intégrables, le théorème central limite implique que la moyenne empirique \bar{X}_n converge p.s. vers $m = \mathbb{E}[X_1]$. Les principes de grandes déviations ont pour objectif d'évaluer la probabilité d'événements rares, du type $\{\bar{X}_n \in A\}$ où $A \subset \mathbb{R}$ ne contient pas m .

Exercice 39 On considère la partie $A = [x, +\infty)$ où $x > m$.

1. Soit $t > 0$ Montrer en utilisant une inégalité de Markov sur $\exp(tS_n)$ que

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp(-n(xt - \phi(t)))$$

où $\phi : \mathbb{R} \rightarrow]-\infty, \infty]$ est la log-Laplace de X_1 :

$$\phi(t) = \log \mathbb{E}[\exp(tX_1)].$$

2. En déduire que

$$\frac{1}{n} \log \mathbb{P}(\bar{X}_n \geq x) \leq -\sup_{t>0} (xt - \phi(t)).$$

On note $I(x) = \sup_{t>0} (xt - \phi(t))$.

3. Dans le cas suivants, expliciter la fonction ϕ puis la fonction I (attention aux valeurs potentiellement infinies qu'elles peuvent prendre) :

- (a) $X_1 \sim \mathcal{B}(p)$,
- (b) $X_1 \sim \mathcal{P}(\lambda)$,
- (c) $X_1 \sim \mathcal{E}(\lambda)$.

L'exercice précédent montre qu'on peut s'attendre à ce que ces probabilités d'événements rares soient majorées par une exponentielle décroissante en n . Le théorème suivant³, très général et s'appliquant même quand X_1 n'est pas intégrable, montre que c'est le cas et fournit également une borne inférieure.

Théorème 3.4.1 (Cramér-Chernov) Soient (X_n) des variables réelles i.i.d, et $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$. Pour $t \in \mathbb{R}$ on note $\phi(t) = \log \mathbb{E}[\exp(tX_1)]$, et pour $x \in \mathbb{R}$,

$$I(x) = \sup_{t \in \mathbb{R}} (xt - \phi(t)).$$

Alors pour tout fermé F de \mathbb{R} ,

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in F) \leq - \inf_{x \in F} I(x)$$

et pour tout ouvert G de \mathbb{R} ,

$$\liminf_n \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in G) \geq - \inf_{x \in G} I(x)$$

Remarque 3.4.2 La fonction I s'appelle fonction de taux ou transformée de Cramér associée à X_1 . Elle est positive, semi-continue inférieurement, convexe.

Exercice 40 Soient (X_n) des variables i.i.d réelles intégrables, de moyenne $m = \mathbb{E}[X_1]$. Dédurre du Théorème 3.4.1 que pour tout $x > m$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{X}_n \geq x) = -I(x).$$

Autrement dit, $\mathbb{P}(\bar{X}_n \geq x) = \exp(-nI(x) + o(n))$

On veut illustrer ce résultat dans le cas $X_1 \sim \mathcal{B}(p)$ où $p \in]0,1[$. On souhaite estimer empiriquement $\frac{1}{n} \log \mathbb{P}(\bar{X}_n \geq x)$ avec $n = 100$ pour diverses valeurs de x entre p et 1. Rappelons dans ce cas la fonction $I(x)$ calculée dans l'exercice 39 :

$$I(x) = x \log \left(\frac{x}{p} \right) + (1-x) \log \left(\frac{1-x}{1-p} \right)$$

pour $x \in [0,1]$, et $+\infty$ ailleurs. Combien d'essais faut-il effectuer pour estimer correctement la probabilité $\mathbb{P}(\bar{X}_n \geq x)$? Tracer la fonction I et dire pourquoi cette probabilité est difficile à estimer empiriquement, à part pour des x très proches de p .

³. Pour une preuve, voir par exemple le livre *Large deviations techniques and applications* de Dembo et Zeitouni

Chapitre 4

Tests et estimateurs classiques

4.1 Estimateurs

Nous commençons par rappeler les définitions générales. Tout au long du chapitre on garde la notation \bar{X}_n pour la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$ d'une suite de variables aléatoires.

4.1.1 Définitions

Définition 4.1.1 *Un modèle paramétrique est une famille de probabilités indexées par un paramètre $\theta \in \Theta$, où $\Theta \subset \mathbb{R}^d$: $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Le modèle est dit identifiable si $\theta \mapsto \mathbb{P}_\theta$ est injective.*

Si X est une variable aléatoire dont la loi appartient à un tel modèle paramétrique \mathcal{P} , une statistique Z est une variable X -mesurable (donc de la forme $\varphi(X)$) ; cette statistique est un estimateur d'un paramètre $g(\theta)$ si presque-sûrement $Z \in g(\Theta)$.

Souvent, le modèle dépendra d'un paramètre n : il s'agira souvent d'un modèle de n réalisations i.i.d. (il sera alors noté $\mathcal{P}^{\otimes n}$ car ses éléments sont de la forme $\mathbb{P}_\theta^{\otimes n}$) mais pas forcément, voir l'exemple ci-après.

Exemple 2

- Si X_1, \dots, X_n est un n -échantillon de loi normale $\mathcal{N}(m, \sigma^2)$ avec m, σ^2 inconnus, alors le modèle naturel associé est

$$\mathcal{P}^{\otimes n} = \{\mathbb{P}_{m, \sigma^2}^{\otimes n} \mid (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$$

où \mathbb{P}_{m, σ^2} est la loi normale $\mathcal{N}(m, \sigma^2)$ sur \mathbb{R} .

- On considère les $n + 1$ premiers pas $(X_k)_{k=0}^n$ d'une chaîne de Markov sur un espace d'état fini E , de loi initiale μ et de matrice de transition P . Le modèle associé est l'ensemble des lois $\mathbb{P}_{\mu, P}^{(n)}$:

$$\mathbb{P}_{\mu, P}^{(n)}(x_0, \dots, x_n) = \mu(x_0)P_{x_0, x_1} \dots P_{x_{n-1}, x_n},$$

qui n'est pas de la forme $\mathcal{P}^{\otimes n}$.

Être un estimateur n'est pas une propriété intéressante : par exemple, n'importe quelle constante de Θ est un estimateur. On va donc définir plusieurs qualités possibles pour un estimateur. On commence par définir le biais et le risque quadratique :

Définition 4.1.2 Soit \mathcal{P} un modèle paramétrique, et Z un estimateur de $g(\theta)$. On suppose que pour tout $\theta \in \Theta$, $\mathbb{E}_\theta(\|Z\|) < \infty$. On définit

$$b(\theta) = \mathbb{E}_\theta(Z) - g(\theta) \quad r(\theta) = \mathbb{E}_\theta((Z - g(\theta))^2)$$

(où, dans les deux cas, \mathbb{E}_θ est l'espérance par rapport à \mathbb{P}_θ) qui sont appelés respectivement le biais de Z et son risque quadratique.

Différentes qualités éventuelles d'un estimateur Z de $g(\theta)$, ou Z_n (lorsque le modèle dépend d'un paramètre n mais que Θ est fixe) seront les suivantes :

- être *sans biais*, c'est-à-dire avoir un biais $b(\theta)$ nul pour tout θ ;
- avoir un risque quadratique faible (mais toute comparaison du risque de deux estimateurs n'a de sens que si elle est vraie pour tout θ) ;
- être *asymptotiquement sans biais*, c'est-à-dire vérifier, $\lim_{n \rightarrow \infty} b_n(\theta) = 0$,
- être *fortement consistant* c'est-à-dire vérifier que pour tout θ on a $Z_n \xrightarrow{\mathbb{P}_\theta} g(\theta)$,
- être (*faiblement*) *consistant*, c'est-à-dire vérifier que pour tout θ on a $Z_n \xrightarrow{\mathbb{P}} g(\theta)$.

Une autre qualité recherchée d'une suite d'estimateurs est l'existence d'une loi asymptotique, c'est-à-dire le fait qu'il existe une suite $(a_n)_n$ avec $a_n \rightarrow \infty$, telle que $a_n(Z_n - g(\theta))$ converge en loi, vers une loi non triviale. On dit dans ce cas que la suite d'estimateurs $(Z_n)_n$ est de *vitesse* $(a_n)_n$. Lorsque $a_n = \sqrt{n}$ et que la loi limite est une normale centrée, on parle de *normalité asymptotique*.

Exercice 41 Soit $\theta > 0$. On considère U_1, \dots, U_n un n -échantillon de loi $\mathcal{U}([0, \theta])$.

1. Montrez que $Z_1 = 2\bar{U}_n$ est un estimateur sans biais, consistant et asymptotiquement normal de θ .
2. Montrez que $Z_2 = \max(U_1, \dots, U_n)$ est un estimateur consistant de θ et que $n(Z_2 - \theta)$ converge en loi vers une loi que l'on identifie.
3. Pour un θ quelconque, simulez un 100-échantillon U_1, \dots, U_n et définissez les estimateurs Z_1 et Z_2 correspondant aux 100 valeurs de n . Tracez les trajectoires de Z_1 et Z_2 . Lequel des deux estimateurs semble converger le plus vite vers θ ?

Exercice 42 Soit (X_1, \dots, X_n) un échantillon de loi admettant un moment d'ordre deux. On propose l'estimateur suivant pour la variance :

$$\hat{s}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Montrer qu'il est fortement consistant mais biaisé. Montrer que $\frac{n}{n-1} \hat{s}_n^2$ est sans biais.

4.1.2 Méthode des moments

Une manière de construire des estimateurs est la méthode des moments. Si $g(\theta)$ est une fonction des moments, on l'estime par la même fonction mais évaluée en les moments empiriques.

Exemple 3 Soit X_1, \dots, X_n un n -échantillon de loi $b(a, b)$ (dont la densité est donnée par $\frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$, où $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$). On a

$$\mathbb{E}(X) = \frac{a}{a+b} \quad \text{var } X = \frac{ab}{(a+b)^2(a+b+1)}.$$

On en tire

$$a = \mathbb{E}(X) \left(\frac{\mathbb{E}(X)(1 - \mathbb{E}(X))}{\mathbb{E}(X^2) - \mathbb{E}(X)^2} - 1 \right) \quad b = (1 - \mathbb{E}(X)) \left(\frac{\mathbb{E}(X)(1 - \mathbb{E}(X))}{\mathbb{E}(X^2) - \mathbb{E}(X)^2} - 1 \right).$$

La méthode des moments suggère donc comme estimateurs

$$\hat{a} = \bar{X}_n \left(\frac{\bar{X}_n(1 - \bar{X}_n)}{\bar{X}_n^2 - \bar{X}_n^2} - 1 \right) \quad \hat{b} = (1 - \bar{X}_n) \left(\frac{\bar{X}_n(1 - \bar{X}_n)}{\bar{X}_n^2 - \bar{X}_n^2} - 1 \right).$$

Ces estimateurs sont fortement consistants d'après la loi des grands nombres.

Exercice 43

- Montrez que Z_1 dans l'exercice 41 ci-dessus aurait pu être trouvé par la méthode des moments.
- Soit P_1, \dots, P_n un n -échantillon de loi de Poisson de paramètre λ . En utilisant les formules pour l'espérance et la variance de cette loi, proposer deux estimateurs de λ différents. Ces estimateurs sont-ils biaisés ? Sont-ils consistants ? Tenter de les comparer par simulation (on pourra se limiter à $\lambda \in \Lambda = [1, 3]$).

4.1.3 Méthode par insertion

La méthode par insertion est similaire : si $g(\theta)$ s'écrit comme une fonction d'un autre paramètre $h(\theta)$, par exemple $g(\theta) = \psi(h(\theta))$, et que l'on connaît un estimateur Z_h de $h(\theta)$, on propose $\psi(Z_h)$ pour estimateur de $g(\theta)$.

Exemple 4 Soit Z_1, \dots, Z_n un n -échantillon de loi $\mathcal{N}(0, \sigma^2)$. Alors $\mathbb{E}(|Z_1|) = \frac{\sigma}{\sqrt{2\pi}}$; on peut donc proposer $\sqrt{2\pi} \left(\frac{|Z_1| + \dots + |Z_n|}{n} \right)$ comme estimateur de σ .

4.1.4 Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est la plus complexe mathématiquement mais aussi la plus universelle, et elle possède souvent de bonnes propriétés. On suppose que toutes les lois \mathbb{P}_θ sont absolument continues par rapport à une mesure commune μ . On note alors f_θ la densité de \mathbb{P}_θ par rapport à μ ; une

réalisation X de loi \mathbb{P}_θ étant donnée, on propose comme estimation de θ la valeur (si elle est unique) de θ qui maximise la vraisemblance $V : \theta \mapsto f_\theta(X)$.

En général on travaillera avec un n -échantillon X_1, \dots, X_n , de sorte que la densité à considérer sera

$$V_n : \theta \mapsto f_\theta(X_1) \dots f_\theta(X_n),$$

Exemple 5 On considère le modèle $\{\mathbb{P}_{m,\sigma^2}^{\otimes n} \mid (m,\sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$. La fonction à maximiser est

$$\theta \mapsto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-m)^2}{2\sigma^2}}.$$

En passant au log, on obtient

$$\log V_n(\theta) = -n \left(\frac{\overline{X^2} - 2m\overline{X} + m^2}{2\sigma^2} + \log \sigma \right) + \text{constante}.$$

Quel que soit σ^2 , le m maximiseur est \overline{X} . En réinjectant ce résultat dans l'expression on trouve que le σ^2 maximiseur est $\overline{X^2} - \overline{X}^2$.

Exercice 44 Dans le modèle de l'exercice 41, quel est l'estimateur du maximum de vraisemblance ?

4.2 Borne de Cramér-Rao et modèles exponentiels

On a vu qu'il était intéressant de minimiser le risque quadratique d'un estimateur. La borne de Cramér-Rao donne une borne inférieure pour ce risque quadratique, montrant que l'on ne peut espérer faire mieux qu'une certaine quantité.

4.2.1 Minoration du risque

Plaçons-nous dans le cadre d'un modèle régulier, c'est à dire que l'on suppose :

- que Θ est un ouvert de \mathbb{R}^d ,
- que toutes les lois \mathbb{P}_θ ont même support et sont absolument continues par rapport à une mesure commune μ , et on note encore $f_\theta = \frac{d\mathbb{P}_\theta}{d\mu}$,
- que $\theta \mapsto \log f_\theta$ est deux fois continûment dérivable en μ -presque tout point, et de carré intégrable.

Définition 4.2.1 On suppose que le modèle $\{\mathbb{P}_\theta \mid \theta \in \Theta \subset \mathbb{R}\}$ est régulier. On appelle information de Fisher du modèle la fonction

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right]$$

Théorème 4.2.2 (borne de Cramér-Rao) Pour $\{\mathbb{P}_\theta \mid \theta \in \Theta\}$ un modèle régulier, on a (sous des hypothèses de régularité supplémentaires) que tout estimateur sans biais de $g(\theta)$ a un risque quadratique qui vérifie

$$r(\theta) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

La preuve est donnée dans *Statistique mathématique en action* de Rivoirard et Stoltz, ou dans votre cours de statistiques de M1.

Exercice 45 Pour une observation X dont l'ensemble des lois possibles est formé par les lois de Poisson de paramètre $\theta \in \mathbb{R}_+^*$, calculer la fonction d'information de Fisher.

On dispose maintenant d'un n -échantillon (X_1, \dots, X_n) de loi $\mathcal{P}(\theta)$. Que devient la fonction $I(\theta)$?

On propose comme estimateur de θ la moyenne empirique \bar{X}_n . Pourquoi est-il sans biais ? Calculer son risque quadratique et le comparer à la borne de Cramér-Rao.

Exercice 46 Dans le cas de l'exercice 41, le théorème précédent s'applique-t-il ? Calculer l'information de Fisher du modèle, puis comparer le risque quadratique des estimateurs Z_1, Z_2 à la borne de Cramér-Rao.

4.3 Intervalles de confiance

Définition 4.3.1 Un intervalle de confiance pour $g(\theta)$ est un intervalle aléatoire $I(\omega)$, dont les bornes sont des fonctions mesurables de X . On dit que l'intervalle de confiance est de niveau (de confiance) $1 - \alpha$ pour $\alpha \in]0, 1[$, si $\mathbb{P}(g(\theta) \in I) \geq 1 - \alpha$.

Un intervalle de confiance asymptotique pour $g(\theta)$ est la donnée pour tout n d'un intervalle de confiance $I_n(\omega)$. On dit qu'il est niveau (de confiance) asymptotique $1 - \alpha$ pour $\alpha \in]0, 1[$, si $\liminf_{n \rightarrow \infty} \mathbb{P}(g(\theta) \in I_n) \geq 1 - \alpha$.

Un intervalle peut être bilatère, c'est-à-dire de la forme $I(\omega) = [A(\omega), B(\omega)]$, ou bien unilatère, c'est-à-dire de la forme $I(\omega) = [A(\omega), +\infty[$ ou bien $I(\omega) =]-\infty, B(\omega)]$. Plus on choisit α petit (donc plus on veut de certitude sur notre estimation), plus l'intervalle devra être grand (et donc moins on aura d'information – mais avec plus de certitude).

Exercice 47 Soit X_1, \dots, X_n un n -échantillon de loi normale de variance connue $\mathcal{N}(m, \sigma_0^2)$. On propose l'estimateur \bar{X}_n de m . Donner un intervalle de confiance pour m en utilisant le fait que $\bar{X}_n - m$ suit une loi connue.

Pour $\beta \in [0, 1]$ et une variable aléatoire réelle X , on appelle quantile d'ordre β de la loi de X la quantité

$$q_\beta = \inf\{x \in \mathbb{R} \mid \mathbb{P}(X \leq x) \geq \beta\}.$$

Supposons qu'une suite d'estimateurs $(Z_n)_n$ est de vitesse $(a_n)_n$, c'est à dire que

$$a_n(Z_n - \theta) \xrightarrow{\mathcal{L}} \text{loi limite.}$$

Alors si l'on note $q_{\alpha/2}$, $q_{1-\alpha/2}$ les quantiles de la loi limite que l'on suppose que ce ne sont pas des atomes¹, on a immédiatement un intervalle de confiance asymptotique de niveau $1 - \alpha$:

$$I_n = \left[Z_n + \frac{q_{\alpha/2}}{a_n}, Z_n + \frac{q_{1-\alpha/2}}{a_n} \right]$$

Exercice 48 On considère le modèle de régression linéaire suivant :

$$Y = \beta_1 f(x) + \beta_2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

où β_1 , β_2 et σ^2 sont des paramètres inconnus à estimer.

1. Donner des estimateurs des paramètres β_1 , β_2 , σ^2 pour un échantillon $(x_1, Y_1), \dots, (x_n, Y_n)$.
Écrire une fonction qui simule le nuage de points correspondant à l'échantillon précédent pour des points x_1, \dots, x_n répartis uniformément sur $[0, 1]$ et la fonction f définie par $f(x) = (1 + x)^2$.
2. Écrire un programme qui trace la courbe de régression et qui donne des intervalles de confiance de niveau $1 - \alpha$ pour les paramètres à estimer.
3. Tracer deux régions de confiance de niveau (au moins) 95% pour le couple (β_1, β_2) : l'une en forme de rectangle en utilisant les intervalles de confiance précédents et l'autre en forme d'ellipse.

Un résultat utile pour discuter de la normalité asymptotique d'estimateurs est ce que l'on appelle la méthode delta :

Lemme 4.3.2 Soit $(\hat{\theta}_n)_n$ une suite de variables aléatoires telle que pour tout $\theta \in \Theta$, $a_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} Z_\theta$ où $(a_n)_n$ est une suite qui croît vers $+\infty$ et Z_θ une variable aléatoire dont la loi dépend de θ . Soit g une fonction à valeurs dans \mathbb{R}^q , différentiable sur un ouvert contenant Θ , de différentielle notée Dg . On a $a_n(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{\mathcal{L}} Dg(\theta)(Z_\theta)$.

Autrement dit, en général, si $(\hat{\theta}_n)_n$ est de vitesse $(a_n)_n$, alors $(g(\hat{\theta}_n))_n$ aussi.

L'exercice suivant utilise la simulation pour estimer le niveau réel d'un intervalle de confiance, c'est-à-dire la valeur de la probabilité $\mathbb{P}(g(\hat{\theta}) \in I)$. Il utilise l'inégalité de Hoeffding :

Proposition 4.3.3 (Inégalité de Hoeffding) Soient X_1, \dots, X_n des variables aléatoires telles que $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ pour tout i . Alors pour tout $t \geq 0$ on a

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq t) \leq 2 \exp \frac{-2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}.$$

1. si $X_n \xrightarrow{\mathcal{L}} X$ alors $\mathbb{P}(X_n \leq x) \rightarrow \mathbb{P}(X \leq x)$ n'est vrai a priori que si x n'est pas un atome de X .

Exercice 49 Soit B_1, \dots, B_n un n -échantillon de loi de Bernoulli $\mathcal{B}(p)$.

1. On propose l'estimateur \bar{B}_n de p . Sur la base de cet estimateur, donnez deux intervalles de confiance (non asymptotiques) pour p , d'abord en utilisant une majoration simple de $p(1-p)$, l'autre par l'inégalité de Hoeffding.
2. Quelle convergence en loi a-t-on pour $\bar{B}_n - p$? Déduire de cette propriété et du lemme de Slutsky un intervalle de confiance asymptotique pour p .
3. Montrer grâce à la méthode delta que pour $g(x) = 2 \arcsin \sqrt{x}$ on a la convergence en loi $\sqrt{n}(g(\bar{B}_n) - g(p)) \rightarrow \mathcal{N}(0,1)$. En déduire un nouvel intervalle de confiance asymptotique pour p .
4. Pour $n = 10, 50, 100$ et différentes valeurs de p , répéter $N = 10000$ fois l'opération suivante : simuler B_1, \dots, B_n , calculer les quatre intervalles ci-dessus et la proportion, sur les N réalisations, de fois où p est bien dans l'intervalle. On est en train d'estimer la probabilité $\mathbb{P}(p \in I)$ à partir d'un N -échantillon de loi binomiale. Quelle précision donne l'expérience ci-dessus ?
5. Estimer le niveau réel des quatre intervalles.

4.4 Tests d'hypothèses : définitions générales

Nous partons d'un modèle paramétrique $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$, et nous donnons deux sous-ensembles disjoints Θ_0 et Θ_1 de Θ . Nous ne supposons pas que $\Theta_0 \cup \Theta_1 = \Theta$. Nous allons considérer les deux hypothèses suivantes :

$$H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1.$$

On verra toujours H_0 comme l'hypothèse *a priori*, et H_1 comme l'hypothèse alternative. Autrement dit, le test vise "à ne rejeter H_0 que si l'on a de bonnes raisons de le faire", et le choix des critères de rejet va dépendre de la forme de H_1 .

Exemple 6 Dupond et Dupont jouent à pile ou face : pile fait gagner Dupond, face fait gagner Dupont. *A priori*, la pile est équilibrée (la probabilité p de faire pile vaut $1/2$) mais les deux policiers se soupçonnent de tricher. Si Dupond veut faire le test, il va naturellement considérer $H_0^d : p = 1/2$ contre $H_1^d : p < 1/2$ (puisque si Dupont triche, c'est pour gagner). Inversement, si Dupont veut faire le test, il va considérer $H_0^i : p = 1/2$ contre $H_1^i : p > 1/2$. On voit bien que les ensembles Θ_0 et Θ_1 ne sont pas complémentaires, ce qui correspond à des hypothèses faites sur le modèle, hypothèses qu'il va s'agir d'exploiter.

Définition 4.4.1 Un test de H_0 contre H_1 est une fonction $\phi(X)$ à valeurs dans $\{0,1\}$, à laquelle on associe la règle de décision : si $\phi(X) = 0$, on conserve H_0 et si $\phi(X) = 1$, on rejette H_0 . On définit les erreurs de première espèce et de seconde espèce associées :

$$\begin{array}{ll} \underline{\alpha} : \Theta_0 & \rightarrow [0,1] \\ \theta & \mapsto \mathbb{P}_\theta(\phi(X) = 1) \end{array} \quad \begin{array}{ll} \underline{\beta} : \Theta_1 & \rightarrow [0,1] \\ \theta & \mapsto \mathbb{P}_\theta(\phi(X) = 0). \end{array}$$

On dit qu'un test est de niveau α si $\sup_{\theta \in \Theta_0} \underline{\alpha} \leq \alpha$.

Les erreurs de première et seconde espèce caractérisent les probabilités des deux manières de se tromper : respectivement, rejeter H_0 à tort (donc observer $\phi(X) = 1$ alors que $\theta \in \Theta_0$), ou conserver H_0 à tort (donc observer $\phi(X) = 0$ alors que $\theta \in \Theta_1$). On appelle *puissance* du test la fonction $1 - \beta$.

Comme pour les intervalles de confiance, il est souvent utile de trouver la loi d'une certaine variable aléatoire sous H_0 , ou au moins une convergence en loi. Dans l'exercice suivant, on utilise le théorème de Cochran (qu'on énonce ici dans un cadre un peu plus restreint que le résultat classique) et la définition de la loi de Student pour établir un tel comportement, et en déduire un test.

Théorème 4.4.2 (Cochran) *Soit X un vecteur gaussien centré réduit à valeurs dans \mathbb{R}^d . Soit une décomposition en sous-espaces orthogonaux $\mathbb{R}^d = \bigoplus_{i=1}^r E_i$, et soit P_1, \dots, P_r les projections orthogonales sur chacun de ces sous-espaces. Alors les projections $(P_i X)_{1 \leq i \leq r}$ sont des vecteurs gaussiens indépendants, centrés, réduits, à valeurs dans E_1, \dots, E_r respectivement.*

En particulier, les normes $(\|P_i X\|^2)_{1 \leq i \leq r}$ sont des v.a. indépendantes de lois respectives $\chi^2(\dim E_1), \dots, \chi^2(\dim E_r)$. Le théorème de Cochran se démontre assez simplement par des décompositions linéaires.

Définition 4.4.3 *Soit X et Y des v.a. indépendantes de lois respectives $\mathcal{N}(0,1)$ et $\chi^2(d)$. Alors la variable aléatoire $Z = \frac{X}{\sqrt{Y/d}}$ suit une loi appelée loi de Student*

à d degrés de liberté, et notée $\mathcal{T}(d)$. Elle a pour densité $\frac{1}{\sqrt{d\pi}} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \left(1 + \frac{x^2}{d}\right)^{-\frac{d+1}{2}}$.

Exercice 50 *Soit X_1, \dots, X_n un n -échantillon de loi $\mathcal{N}(m, \sigma^2)$ avec m et σ inconnus. On cherche à définir un test de $H_0 : m = m_0$ contre $H_1 : m > m_0$.*

- 1. Soit \bar{X}_n et \hat{s}_n^2 la moyenne empirique et la variance empirique de (X_1, \dots, X_n) . Montrer que sous l'hypothèse H_0 , la variable aléatoire $\sqrt{n-1} \frac{\bar{X}_n - m_0}{\sqrt{\hat{s}_n^2}}$ suit la loi $\mathcal{T}(n-1)$.*
- 2. En déduire un test de niveau α . Justifier que de plus, sa puissance tend vers 1 quand $n \rightarrow \infty$.*
- 3. Définir une fonction qui, si on lui donne un n -échantillon X_1, \dots, X_n en entrée, donne le résultat du test.*
- 4. On se place dans le cas $m = m_0 = 2$, $\alpha = 5\%$, et, pour conserver notre ignorance de σ , on tirera au hasard une valeur dans $[1, 2]$; on veut estimer le niveau réel $\underline{\alpha}(m_0)$. On va donc faire N fois l'expérience qui consiste à simuler n variables de loi $\mathcal{N}(m, \sigma^2)$, à appliquer le test à ce n -échantillon et à compter combien de fois, sur les N , on rejette (à tort) l'hypothèse H_0 . Sachant que la vraie valeur de $p = \underline{\alpha}(m_0)$ est de l'ordre de 0,05, quelle valeur de N choisir pour avoir une estimation (de niveau de confiance 95%) de p à 0,01 près ?*
- 5. Implémenter l'expérience ci-dessus, et estimer le niveau réel du test pour différentes valeurs de la taille n par exemple (5,10,50,100).*

6. On veut maintenant estimer la puissance du test. Estimer cette quantité pour $n = 5, 10, 50, 100$ et m variant de 2,1 à 3 par pas de 0,1. Représenter les résultats.

4.5 Tests du chi-deux

Le test du χ^2 est sans doute l'un des tests les plus connus et les plus courants. Il permet de tester si un échantillon suit bien une loi donnée (test d'ajustement à une loi), ou si elle appartient à une famille de lois (test d'ajustement à une famille), ou encore si des variables sont indépendantes, suivent la même loi inconnue a priori, etc.

4.5.1 Ajustement à une loi

On dispose d'un n -échantillon (X_1, \dots, X_n) i.i.d. dont les variables sont à valeurs dans un ensemble fini $E = \{u_1, \dots, u_d\}$. On formule l'hypothèse H_0 que ces variables suivent une loi fixée $p^{\text{ref}} = (p_j^{\text{ref}})_{1 \leq j \leq d}$. Ainsi

$$\begin{aligned} H_0 : X_i &\sim p^{\text{ref}}, \\ H_1 : X_i &\not\sim p^{\text{ref}}. \end{aligned}$$

Pour cela, on compare la loi empirique des $(X_i)_{1 \leq i \leq n}$ avec la loi de référence, en utilisant la méthode des moments. Notons la fréquence empirique

$$\hat{p}_{j,n} = \frac{\sum_{i=1}^n \mathbb{1}_{X_i=u_j}}{n}.$$

Alors on considère la pseudo-distance suivante, appelée *statistique de Pearson* :

$$D_n^2 = n \sum_{j=1}^d \frac{(\hat{p}_{j,n} - p_j^{\text{ref}})^2}{p_j^{\text{ref}}}.$$

On rappelle que la loi du χ^2 à k degrés de liberté, notée $\chi^2(k)$, est la loi de $N_1^2 + \dots + N_k^2$ où les N_i suivent des lois normales centrées réduites indépendantes. Sa densité est $\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \mathbb{1}_{x \geq 0}$. Une preuve du résultat suivant est donnée dans *Statistique mathématique en action* de Rivoirard et Stoltz; elle repose sur un théorème central limite multi-dimensionnel, et sur le théorème de Cochran.

Théorème 4.5.1 Sous H_0 , $D_n^2 \xrightarrow{\mathcal{L}} \chi^2(d-1)$. Sous H_1 , $D_n^2 \xrightarrow{p.s.} +\infty$.

Par conséquent on définit le test

$$\phi(X_1, \dots, X_n) = \mathbb{1}_{D_n^2 > c_{d-1, 1-\alpha}}$$

où $c_{d-1, 1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(d-1)$. Ce test est donc de niveau asymptotique α .

Exercice 51 Test du chi-deux d'ajustement. Un ordinateur possède un générateur pseudo-aléatoire de nombres choisis au hasard dans l'ensemble des entiers de 0 à 9. On dispose d'un échantillon de taille $N = 1000$ de chiffres tirés par ce générateur. Les résultats sont répartis dans le tableau suivant.

Chiffres	0	1	2	3	4	5	6	7	8	9
Observations	120	87	115	103	91	109	92	112	94	77

1. On veut tester l'hypothèse d'équiprobabilité pour chaque chiffre. Mettre en œuvre le test du χ^2 . Choisissez vous d'accepter l'hypothèse d'équiprobabilité pour l'échantillon précédent, et si oui pour quel niveau α ?

Le plus grand α pour lequel on conserve H_0 est parfois appelé p-valeur ; c'est une variable aléatoire.

2. Faites de même en remplaçant la table précédente par une table générée via la fonction `random` de Python.

4.5.2 Ajustement à une famille de lois

On reprend la même situation, mais cette fois-ci on fixe une famille de lois de probabilités sur E , notée $\mathcal{P} = \{p(\theta), \theta \in \Theta\}$ où Θ est un ouvert de \mathbb{R}^k , avec $k < d - 1$. On teste donc $H_0 : \mathcal{L}(X_i) \in \mathcal{P}$ contre $H_1 : \mathcal{L}(X_i) \notin \mathcal{P}$.

Pour ce faire, on commence par construire un estimateur $\hat{\theta}_n$ de θ par la méthode du maximum de vraisemblance. On en déduit la loi $p(\hat{\theta}_n) = (p_j(\hat{\theta}_n))_{1 \leq j \leq d}$.

On construit alors la statistique de Pearson associée à cette loi $p(\hat{\theta}_n)$:

$$\hat{D}_n^2 = n \sum_{j=1}^d \frac{(\hat{p}_{j,n} - p_j(\hat{\theta}_n))^2}{p_j(\hat{\theta}_n)}.$$

Théorème 4.5.2 Si l'application $p : \theta \mapsto p(\theta) = (p_j(\theta))_{1 \leq j \leq d}$ est injective, de classe C^2 , qu'aucune de ses composantes ne s'annulent sur Θ , et que ses k dérivées partielles sont linéairement indépendantes en tout point de Θ , alors sous H_0 ,

$$\hat{D}_n^2 \xrightarrow{\mathcal{L}} \chi^2(d - 1 - k).$$

De plus, sous H_1 , si en plus $d(\mathcal{L}(X_i), \mathcal{P}) > 0$, alors $\hat{D}_n^2 \xrightarrow{P-s.} +\infty$.

Exercice 52 On considère n couples d'observations $(X_1, Y_1), \dots, (X_n, Y_n)$, où les X_i et Y_i sont toutes à valeurs dans $E = \{u_1, \dots, u_d\}$. On suppose de plus qu'elles chargent tous les points de E . On veut tester l'hypothèse H_0 : les X_i sont indépendantes des Y_i , contre H_1 l'hypothèse contraire.

Écrire ce problème comme un problème d'ajustement à une famille de lois sur E^2 . Exprimer l'estimateur du maximum de vraisemblance, puis la statistique

de Pearson associée, en fonction des fréquences empiriques :

$$\begin{aligned}\hat{p}_{j_1,n} &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i=u_{j_1}}, \\ \hat{q}_{j_2,n} &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i=u_{j_2}}, \\ \hat{r}_{j_1,j_2,n} &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(X_i,Y_i)=(u_{j_1},u_{j_2})}.\end{aligned}$$

En déduire un test de niveau asymptotique α . C'est le test du χ^2 d'indépendance.

Exercice 53 Cette fois-ci on dispose de deux échantillons, (A_1, \dots, A_n) et (B_1, \dots, B_m) , toujours à valeurs dans E . Notons α la loi des A_i et β la loi des B_i . On cherche à tester l'hypothèse $H_0 : \alpha = \beta$ contre $H_1 : \alpha \neq \beta$.

Montrer que cela revient à un test d'indépendance sur une permutation aléatoire des $n+m$ couples d'observations $(A_1,1), \dots, (A_n,1), (B_1,2), \dots, (B_m,2)$, et donc qu'on peut se ramener à l'exercice précédent.

C'est le test du χ^2 d'homogénéité.

4.6 Test de Kolmogorov et dérivés

Une autre manière de construire des tests d'ajustement à une loi donnée est d'exploiter les Théorèmes 2.3.3 et 2.3.4 sur les fonctions de répartition. On explore cette idée par un exercice.

Exercice 54 Tests de Kolmogorov et de Lilliefors. Soit X_1, X_2, \dots une suite de v.a. i.i.d, de fonction de répartition F . On suppose que F est continue. Pour tout n , on définit la fonction de répartition empirique de l'échantillon (X_1, \dots, X_n) par

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \leq x}.$$

On définit la distance de Kolmogorov-Smirnov

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|.$$

On rappelle le Théorème 2.3.4 : $\sqrt{n} D_n$ converge en loi vers une loi dite de Kolmogorov-Smirnov, qui ne dépend pas de la loi des X_k .

1. Soient $(X^{(1)}, \dots, X^{(n)})$ les valeurs de l'échantillon dans l'ordre croissant. Montrer que

$$D_n = \max \left(\max_{k=1, \dots, n} \{k/n - F(X^{(k)})\}, \max_{k=1, \dots, n} \{F(X^{(k)}) - (k-1)/n\} \right).$$

Ecrire une fonction Dn qui calcule cette quantité pour un échantillon et une fonction F données.

2. Illustrer la convergence presque sûre vers 0 de D_n .
3. Illustrer la convergence en loi de $\sqrt{n}D_n$ vers une limite K . La loi de K est implémentée dans `scipy.stats` sous le nom `scs.kstwobign`.
4. En s'inspirant de cette convergence en loi et du test du χ^2 , proposer un test d'ajustement de la loi de l'échantillon à une loi F_0 donnée : quelles sont les hypothèses H_0 et H_1 du test, quand rejetez-vous H_0 ?
Ce test est appelé test de Kolmogorov.
5. Mettez en oeuvre le test, au niveau asymptotique $\alpha = 5\%$, dans les cas suivants :
 - échantillon de loi normale standard avec F_0 une loi normale standard,
 - échantillon de loi exponentielle avec F_0 une loi exponentielle,
 - échantillon de loi normale standard avec F_0 une loi de Laplace.
6. Le test de Lilliefors est un test de normalité basé sur le test de Kolmogorov-Smirnov. On considère la statistique

$$L_n = \sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{\Phi}_n(x)|,$$

où $\hat{\Phi}_n$ est la fonction de répartition de la loi $\mathcal{N}(\bar{X}, S^2)$, avec

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k, \quad S^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2.$$

Illustrer le fait que L_n converge en loi quand $n \rightarrow \infty$ si les X_i suivent une loi normale, et comparer la limite L avec K . Prouver que L_n tend presque-sûrement vers l'infini si les X_i ne suivent pas une loi normale.

7. Estimez le 95%-quantile de L et utilisez-le pour proposer un test de normalité de niveau asymptotique 5%.

Le test de Kolmogorov peut également être modifié en un test d'homogénéité appelé *test de Kolmogorov-Smirnov* : on dispose de deux échantillons, $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_m)$, et on veut savoir s'ils ont la même loi, sans chercher à connaître cette loi. Pour cela, on note \hat{F}_n (resp. \hat{G}_m) leur fonction de répartition empirique respective, et on calcule $D_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|$. Un résultat analogue au Théorème 2.3.4 assure que $\sqrt{\frac{nm}{n+m}} D_{n,m}$ converge en loi vers une loi universelle lorsque $n, m \rightarrow \infty$. Il s'agit ensuite d'adapter le test précédent à cette loi.

4.7 Exercice supplémentaire

Exercice 55 Etude de la robustesse² d'un test. Soit une suite d'observations X_1, X_2, \dots, X_n i.i.d. de loi μ . On note \bar{X} la moyenne empirique et S^2 la

2. La robustesse d'un test est définie comme la non-sensibilité de la procédure de test à la loi des observations. Le test asymptotique sur la moyenne fondé sur le TCL est ainsi robuste sur l'ensemble des lois admettant un moment d'ordre deux.

variance empirique (version biaisée) :

$$\bar{X} = \frac{1}{n} \sum_{t=1}^n X_t, \quad S^2 = \frac{1}{n} \sum_{t=1}^n (X_t - \bar{X})^2.$$

1. On suppose que μ est une loi gaussienne, de moyenne m et de variance σ^2 . Quelle est la loi de S^2 ?
2. On pose $H_0 : \sigma^2 \leq 1$. Dédurre du résultat précédent un test de niveau (exactement) α .

On se demande maintenant si le test construit précédemment s'étend à des lois non-gaussiennes, c'est-à-dire, si lorsque μ n'est plus gaussienne, le test précédent (avec la même statistique de test, avec le même choix pour la zone de rejet) est toujours, au moins asymptotiquement, de niveau α . On admet pour l'instant les résultats suivants :

- si $c_{n-1,\alpha}$ désigne le α -quantile de la loi du χ^2 à $n-1$ degrés de liberté, alors

$$\frac{c_{n-1,\alpha} - n}{\sqrt{2}\sqrt{n}} \rightarrow u_\alpha,$$

où u_α désigne le α -quantile de la loi gaussienne standard ;

- on a la convergence en loi suivante :

$$\sqrt{n} \left(\frac{S^2}{\sigma^2} - 1 \right) \rightsquigarrow \mathcal{N}(0, \kappa - 1),$$

où κ désigne la kurtosis de μ , définie par

$$\kappa = \frac{\mu_4}{\mu_2^2},$$

avec, pour $k \in \mathbb{N}$, $\mu_k = \mathbb{E} \left[(X - \mathbb{E}[X])^k \right]$ (notez qu'en particulier, $\mu_2 = \sigma^2$ est la variance de μ .)

3. On note Φ la fonction de répartition de la loi gaussienne. Prouver qu'asymptotiquement, le test précédent est de niveau $1 - \Phi(u_\alpha \sqrt{2}/\sqrt{\kappa - 1})$, c'est-à-dire que

$$\limsup_n \sup_{\sigma^2 \leq 1} \mathbb{P}_{\sigma^2}(\text{rejet du test}) \leq 1 - \Phi \left(\frac{u_\alpha \sqrt{2}}{\sqrt{\kappa - 1}} \right).$$

On a donc prouvé la non-robustesse contre une modification de la valeur de la kurtosis.

4. Mettons en évidence cette non-robustesse par voie de simulations. Pour $n = 20$, estimer par méthode de Monte-Carlo le niveau réel du test proposé à la question (2), pour $\alpha = 5\%$ et μ donné, d'une part par une loi de Laplace, et d'autre part par un mélange de gaussiennes $0,95\mathcal{N}(0, 1) + 0,05\mathcal{N}(0, 9)$ (il faut renormaliser ces deux lois pour être dans l'hypothèse nulle).

5. Il faut encore prouver les deux résultats que nous avons admis. Ils découlent tous deux de la convergence en loi suivante, à démontrer (on note m l'espérance de μ) :

$$\sqrt{n}(S^2 - \sigma^2) = \sqrt{n} \left(\frac{1}{n} \sum_{t=1}^n (X_t - m)^2 - \sigma^2 \right) - \sqrt{n}(\bar{X}_n - m)^2$$

converge en loi vers une $\mathcal{N}(0, \mu_4 - \mu_2^2)$.

Chapitre 5

Chaînes de Markov

Les chaînes de Markov sont des objets couramment utilisés en simulation, parce qu'elles apparaissent dans de nombreux problèmes de modélisation. Un autre intérêt est l'estimation de quantités du type $\int f d\pi$, où π est une mesure de probabilité que l'on ne sait pas bien simuler. Si l'on arrive à construire une chaîne de Markov de probabilité invariante π , pour lesquels les théorèmes classiques de convergence s'appliquent, alors la distribution au temps long de la chaîne sera proche de μ .

Il est recommandé de revoir au préalable vos cours sur la théorie générale des chaînes de Markov, même si nous nous concentrerons sur le cas plus simple des chaînes à espace d'états finis.

5.1 Simulation et résultats classiques

Les notations utilisées dans l'ensemble du texte seront les suivantes : $(X_n)_n$ sera toujours une chaîne de Markov de matrice de transition (ou *noyau* de transition) $(P(x,y))_{x,y \in E}$ sur un ensemble E de cardinal fini d , avec la convention $P(x,y) = \mathbb{P}(X_{n+1} = y | X_n = x)$.

Si $\mu_0 = (\mu_0(x))_{x \in E}$ (vu comme un vecteur ligne) est la loi de X_0 alors $\mu_n = \mu_0 P^n$ est la loi de X_n . De même, si f est une fonction sur E (vue comme un vecteur colonne) alors Pf sera la fonction

$$Pf(x) = \sum_{y \in E} P(x,y)f(y) = \mathbb{E}_x(f(X_1)).$$

On notera \mathbb{P}_x (resp. \mathbb{P}_μ) les probabilités conditionnelles à $X_0 = x$ p.s. (resp. sous l'hypothèse que X_0 suit la loi μ), de même pour \mathbb{E}_x (resp. \mathbb{E}_μ).

5.1.1 Trajectoire

Supposons que la matrice de transition est donnée sous la forme d'une liste de listes, par exemple $P = [[1/2, 1/3, 1/6], [1/3, 1/3, 1/3], [1/2, 0, 1/2]]$ pour la matrice $P = \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \end{pmatrix}$; notons que c'est bien la forme qu'aura P si on la code par une variable `np.array`, avec par exemple

$$P = \text{np.array}([[1/2, 1/3, 1/6], [1/3, 1/3, 1/3], [1/2, 0, 1/2]]).$$

Dans ce cas, et si l'on indexe (suivant l'habitude de Python) les trois sommets comme 0, 1 et 2 alors conditionnellement à $X_n = i$ la variable X_{n+1} vaut 0, 1 ou 2 avec des probabilités données par $P[i]$. On peut alors utiliser la commande `rnd.choice(d,p=P[i])` où d est le cardinal de l'espace d'état (ici $d = 3$).

Exercice 56 *Écrivez une fonction Python qui prend en entrée X_0 , n et P et simule une trajectoire (X_0, \dots, X_n) .*

Il ressort de l'exercice précédent qu'une chaîne de Markov peut également être donnée sous la forme $X_{n+1} = f_n(X_n, U_{n+1})$ où $(U_n)_n$ est une suite i.i.d. de loi $\mathcal{U}([0,1])$ indépendante de X_0 .

5.1.2 Irréductibilité

Pour une chaîne de Markov à espace d'états finis, il n'y a pas d'état récurrent nul. La chaîne de Markov est dite irréductible si pour tous $x, y \in E$, il existe $n \in \mathbb{N}$ tel que $P^n(x, y) > 0$. Si ce n'est pas le cas, on peut partitionner E en une union disjointe de classes, telles que la restriction de P à chacune de ces classes est irréductible; on parle de *classes irréductibles*; chacune de ces classes est soit entièrement récurrente soit entièrement transitoire. De plus, la chaîne est presque sûrement capturée par une classe récurrente. La plupart du temps, en étudiant un peu la chaîne a priori, on pourra se restreindre à une classe récurrente. Ainsi, les hypothèses de "chaînes de Markov irréductibles récurrentes" ci-après ne coûtent pas très cher dans ce cas.

5.1.3 Période

Supposons la chaîne irréductible. Sa *période* est alors

$$d = \text{pgcd}\{n > 0 \mid P^n(x, x) > 0\}$$

et ne dépend pas de l'état $x \in E$. La chaîne est dite *apériodique* si $d = 1$. C'est par exemple le cas si $P(x, x) > 0$ pour un certain $x \in E$.

Remarque 5.1.1 (Apériodicité par perturbation) *Soit $p \in]0, 1[$, alors la matrice*

$$P_p = (1 - p)P + pI$$

est apériodique, reste irréductible si P l'était, et possède les mêmes mesures invariantes. Pour la trajectoire de la chaîne, cela revient à choisir avec probabilité p de ne pas bouger, et ce à chaque étape.

5.1.4 Mesure invariante et théorème ergodique

Si la chaîne est irréductible avec un espace d'états finis, elle est nécessairement récurrente positive et on sait que dans ce cas il existe une unique mesure de probabilité invariante π .

Exercice 57 *Écrire une fonction Python qui prend en entrée P que l'on suppose irréductible, et donne la mesure invariante π de cette chaîne. On utilisera `np.linalg.eig`, dont on pourra consulter le fichier d'aide, et `np.where`. Ne pas*

oublier que la probabilité invariante est vecteur propre de la transposée de P et pas de P même, et que la sortie doit être une probabilité, donc un vecteur à coefficients positifs et dont la somme vaut 1.

Il existe un résultat de type “loi des grands nombres” pour les chaînes de Markov : c’est le théorème ergodique.

Théorème 5.1.2 *Soit $(X_n)_n$ une chaîne de Markov sur un espace d’état dénombrable E . Si cette chaîne est irréductible et admet une probabilité invariante π , alors pour toute fonction f sur E intégrable par rapport à π , on a*

$$\frac{1}{n} \sum_{k=1}^n f(X_k) \xrightarrow{p.s.} \int f d\pi.$$

Notons que l’irréductibilité assure que la probabilité invariante est unique si elle existe, et qu’elle existe toujours si E est fini. Ce résultat est démontré par exemple dans *Promenade aléatoire* de Benaïm et El Karoui, ou dans *Modélisation stochastique* de Bercu et Chafaï.

Notons également que ce théorème ne requiert pas d’apériodicité. Comme on effectue une moyenne temporelle, les effets de périodicités sont “gommés”.

Il existe également un résultat de type “théorème central de la limite” pour les chaînes de Markov :

Théorème 5.1.3 *Soit $(X_n)_n$ une chaîne de Markov sur un espace d’état fini E . Si cette chaîne est irréductible et qu’on note π sa probabilité invariante, alors pour toute fonction f sur D , il existe σ_f^2 tel que*

$$\sqrt{n} \left(\frac{1}{n} \sum_{k=1}^n f(X_k) - \int f d\pi \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_f^2).$$

Le gros défaut pratique de cet énoncé est que σ_f n’est défini que de manière implicite, de sorte que ce résultat ne peut servir pour donner des intervalles de confiance.

Exercice 58 *Illustrez les deux théorèmes ci-dessus dans le cas de la chaîne de matrice de transition P ci-dessus et pour $f = \mathbf{1}_{\{0\}}$. La variance σ_f étant inconnue, on l’estimera par la variance empirique d’un N -échantillon de X_n pour n assez grand.*

5.1.5 Convergence en loi vers l’équilibre

Le Théorème 5.1.2 ne nous dit pas si X_n converge effectivement en loi vers la mesure invariante π . Et de fait ce n’est pas forcément le cas : imaginons que E est l’union disjointe de deux sous-ensemble E_1 et E_2 , et que la chaîne saute à chaque étape d’un sous-ensemble à l’autre. Alors pour une mesure de départ $\mu_0 = \delta_x$ où $x \in E_1$, on aura μ_n portée par E_1 pour les n pairs et par E_2 pour

les n impairs ; elle ne peut donc pas converger faiblement¹. C'est pourquoi le théorème suivant a des hypothèses plus fortes.

Théorème 5.1.4 *Soit $(X_n)_n$ une chaîne de Markov sur un espace d'état dénombrable E . Si cette chaîne est irréductible, récurrente positive et apériodique et qu'on note π sa probabilité invariante, alors quelle que soit la loi de X_0 ,*

$$X_n \xrightarrow{\mathcal{L}} \pi.$$

Dans le cas d'un espace d'états fini, c'est une conséquence du théorème de Perron-Frobenius.

5.2 Méthodes de Monte-Carlo

Le principe général de la méthode de Monte-Carlo est le suivant : pour toute fonction f sur E , le théorème ergodique donne (sous ses hypothèses) :

$$\frac{1}{n} \sum_{k=1}^n f(X_k) \xrightarrow[n \rightarrow \infty]{\text{p.s.}} \int f d\pi. \quad (5.1)$$

On peut donc espérer calculer $\int f d\pi$ en simulant une trajectoire d'une chaîne de Markov de probabilité invariante π . Notons que c'est différent, que ce soit conceptuellement ou numériquement, d'une estimation de cette intégrale $\mathbb{E}(f(X))$ par une moyenne empirique $\frac{1}{n} \sum_{k=1}^N f(X^{(k)})$ pour $X^{(1)}, \dots, X^{(N)}$ un N -échantillon de X :

- dans l'estimation par la moyenne empirique, on moyenne sur les valeurs de plusieurs réalisations indépendantes au même temps (on a plusieurs ω , un seul temps) et on a besoin de savoir simuler X de loi π ;
- dans l'estimation basée sur (5.1), on moyenne sur les valeurs à différents "temps" d'une seule trajectoire (on a un seul ω , plusieurs temps) et on a besoin de savoir simuler une chaîne de Markov de probabilité invariante π .

Si l'on sait simuler *simplement* les transitions d'une chaîne de Markov de probabilité invariante π , alors la deuxième méthode est plus efficace. Notons qu'elle ne sera intéressante, en particulier, que s'il est moins coûteux numériquement de simuler la chaîne que de calculer tous les $\pi(x)$ et de sommer les $f(x)\pi(x)$, donc plutôt quand on a une description simple de la dynamique de la chaîne, mais que cette chaîne vit dans un espace d'état de grande taille. Mais est-ce si simple de construire une chaîne de Markov (irréductible) de probabilité invariante π ? Nous allons décrire dans la section 5.3 une méthode, dite de Metropolis-Hastings, de simulation d'une chaîne de Markov de probabilité invariante π .

On ne se pose pas ici la question de la vitesse de convergence, que l'on pourra traiter par une étude du spectre de la matrice de transition P : puisque lorsque P est irréductible les constantes sont son seul invariant, on peut (sous certaines hypothèses) estimer $|P^n f(x) - \int f d\pi|$ en fonction de la valeur propre de P

1. On a en fait une convergence de type Césaro.

qui a le module le plus grand après 1 (1 est toujours valeur propre, et c'est la plus grande). Nous ne décrivons pas en détail cette étude spectrale : de toute façon, la méthode se heurte à des calculs pénibles dès que l'on s'intéresse à des problèmes concrets. Nous exploiterons cependant l'idée que l'on peut contrôler la vitesse de convergence pour obtenir un algorithme stochastique de recherche de minima, algorithme appelé le *recuit simulé*, présenté dans la section 5.5.

5.3 Algorithme de Metropolis–Hastings

Dans cette section, nous allons donner une méthode permettant de construire une matrice de transition P qui sera réversible pour une matrice donnée π .

Rappelons la définition de la réversibilité :

Définition 5.3.1 *On dit que la matrice de transition $(P(x,y))_{x,y}$ est réversible par rapport à π si pour tous x,y de E on a*

$$P(x,y) \pi(x) = P(y,x) \pi(y). \quad (5.2)$$

Si P est réversible par rapport à π , alors π est invariante pour P . Plus précisément, l'invariance de π est caractérisée par l'égalité

$$\sum_{y \in E} P(x,y) \pi(x) = \sum_{y \in E} P(y,x) \pi(y),$$

donc l'invariance de π par P correspond à l'égalité "en moyenne" de $P_{x,y} \pi(x)$ alors que la réversibilité correspond à une égalité terme à terme. Ceci explique que (5.2) soit aussi appelée la condition de *bilan détaillé*.

On part d'une probabilité π dont on suppose qu'elle charge tous les points, c'est-à-dire que $\pi(x) > 0$ pour tout $x \in E$, et on suppose donnée une matrice de transition Q , dite matrice de sélection, qui a la propriété $Q(x,y) = 0 \Rightarrow Q(y,x) = 0$. On pose

$$\alpha(x,y) = \min \left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)} \right) \quad (5.3)$$

avec la convention que $\alpha(x,y) = 0$ si $Q(x,y) = 0$.

Lemme 5.3.2 *On définit*

$$P(x,y) = \begin{cases} \alpha(x,y) Q(x,y) & \text{si } x \neq y, \\ 1 - \sum_y \mathbf{1}_{y \neq x} P(x,y) & \text{sinon.} \end{cases}$$

Alors P est un noyau de transition qui est réversible pour π , et irréductible si Q l'est.

La matrice de transition P donnée ci-dessus correspond à un algorithme d'évolution très simple : si l'on a $X_n = x$, alors le choix de la valeur de X_{n+1} se fait de la manière suivante :

1. on choisit y suivant la loi $(Q(x,y))_{y \in E}$,

2. on calcule $\alpha(x,y)$;
3. on tire un U de loi $\mathcal{U}([0,1])$;
 - si $U \leq \alpha(x,y)$ alors on accepte la sélection et on choisit $X_{n+1} = y$;
 - sinon on refuse la sélection et on choisit $X_{n+1} = x$.

Exercice 59 *Prouvez le Lemme 5.3.2.*

Remarque 5.3.3 *Il existe d'autres choix de fonctions α avec les mêmes propriétés, par exemple*

$$\alpha(x,y) = \frac{\pi(y)Q(y,x)}{\pi(y)Q(y,x) + \pi(x)Q(x,y)} \quad (5.4)$$

et $\alpha(x,y) = 0$ si $Q(x,y) = 0$. Ce choix a pour avantage d'être toujours apériodique, mais cela est aussi très souvent le cas pour le choix (5.3) (par exemple dès que $\alpha(x,y) < 1$ pour certains x,y).

Cet algorithme simple permet donc de simuler une chaîne de Markov qui va être réversible par rapport à π . On verra que l'estimation des vitesses de convergence est plus simple pour les chaînes réversibles.

Remarque 5.3.4 *On n'a pas besoin d'explicitier la matrice Q : il suffit de savoir définir un algorithme qui simule la chaîne de transitions données par Q , et de connaître les rapports $Q(x,y)/Q(y,x)$. On n'a pas besoin non plus de connaître explicitement $\pi(x)$: il suffit de connaître les rapports $\pi(x)/\pi(y)$.*

Exercice 60 *Soit $V = (\mathbb{Z}/r\mathbb{Z})^d$ un réseau d -dimensionnel. On appelle configuration de sphères dures sur S une application $x : V \rightarrow \{0,1\}$ telle que $x(v) \neq x(v')$ si v et v' sont voisins. La situation $x(v) = 1$ décrit la présence en v d'une "sphère" qui est assez grosse pour empêcher la présence d'une autre sphère sur les sites voisins (mais pas sur les sites plus lointains). On note M l'ensemble des configurations de sphères dures sur V , et π la probabilité uniforme sur M . Le but est de simuler cette distribution π , ce qui n'est pas évident a priori.*

On considère la chaîne de Markov suivante : si $X_n = x \in M$, alors

1. on choisit v uniformément dans V ,
2. si le site v est libre et que les sites voisins ne sont pas tous libres, on ne fait rien,
3. si le site v est occupé, ou qu'il est libre et que tous les sites voisins de v sont libres, on choisit $x(v) = 0$ ou 1 suivant une $\mathcal{B}(\frac{1}{2})$;

ceci définit la configuration X_{n+1} .

Montrez que c'est l'algorithme de Metropolis pour la fonction (5.4), et définit donc une chaîne de Markov irréductible apériodique de mesure invariante π .

En déduire une estimation numérique du nombre moyen de sphères $\mathbb{E}_\pi(\sum_{v \in V} x(v))$ (en prenant $n = 10\,000$ pour $r = 10$, $d = 2$).

5.4 Mesures de Gibbs

Soit V une fonction de E dans \mathbb{R} , et $T > 0$ un paramètre appelé la température. On appelle mesure de Gibbs (ou de Boltzmann–Gibbs) associée à la fonction V , à la température T , la probabilité définie par

$$\pi_{V,T}(x) = \frac{1}{Z} e^{-V(x)/T}, \quad (5.5)$$

où Z est une constante de normalisation :

$$Z = \sum_{x \in E} e^{-V(x)/T}.$$

Les mesures de Gibbs sont fondamentales en physique, et plus particulièrement en physique statistique où E représente l'espace d'état d'un système physique, $x \in E$ une configuration donnée et $V(x)$ l'énergie de cette configuration. Les mesures de Gibbs sont alors les lois d'équilibre macroscopique du système², et il est important de pouvoir les simuler.

L'une des difficultés de la simulation directe des lois de Gibbs réside dans la constante Z , *a priori* inconnue et dont le calcul est long – ou en tout cas aussi compliqué que le calcul de $\int f d\pi_{V,T}$ – dès que E est de grande taille. C'est là que la Remarque 5.3.4 prend tout son sens. L'algorithme est de plus simplifié par le fait que pour tous x, y de E ,

$$\frac{\pi_{V,T}(y)}{\pi_{V,T}(x)} = e^{-(V(y)-V(x))/T}$$

qui ne dépend que des différences $V(y) - V(x)$, en général simples à calculer. Explicitons dans ce cadre l'algorithme de Metropolis : partant de $x \in E$, l'évolution est donnée simplement par

1. on choisit $y \in E$ suivant la loi $(Q(x,y))_{y \in E}$;
2. on calcule $\alpha(x,y) = \min\left(1, \frac{Q(y,x)}{Q(x,y)} e^{-\Delta V(x,y)/T}\right)$ où $\Delta V(x,y) = V(y) - V(x)$ (ou bien, pour la fonction (5.4), $\alpha(x,y) = \frac{1}{1 + \frac{Q(x,y)}{Q(y,x)} e^{-\Delta V(x,y)/T}}$) ;
3. on tire U de loi $\mathcal{U}([0,1])$;
 - si $U \leq \alpha(x,y)$ alors on accepte la sélection et on choisit $X_{n+1} = y$;
 - sinon on refuse la sélection et on choisit $X_{n+1} = x$.

Le principe de cet algorithme sera plus naturel dans le cas où $Q(x,y) = Q(y,x)$, comme ce sera le cas dans l'exercice 61 ci-dessous.

Notre premier exercice de vraie simulation concerne l'utilisation de la méthode de Metropolis–Hastings non plus pour estimer une intégrale $\int f d\pi$ mais pour simuler π (on va donc utiliser le Théorème 5.1.4) dans un modèle simple d'aimant appelé modèle d'Ising.

2. Au sens où les mesures $\pi_{V,T}$ sont les mesures π qui maximisent l'entropie $S(\pi) = -\int \log \pi(x) d\pi$ sous la contrainte que l'énergie totale $\int V d\pi$ est fixée.

Exercice 61 On considère un réseau fini $R = \{1, \dots, r\}^2$, qui représente les positions des atomes dans un bloc de métal. On note $a \sim b$ si deux sites a et b sont voisins. Chaque atome a a un moment magnétique (une “micro-aimantation”) qui est orienté soit vers le haut, soit vers le bas : on note $\sigma(a) \in \{-1, +1\}$ le moment magnétique (que l’on appelle habituellement “spin”) en $a = (i, j) \in R$. La configuration du bloc est donc décrite par $\sigma = (\sigma(a))_{a \in R}$ et donc l’espace d’états est $E = \{-1, +1\}^R$.

Pour des raisons physiques, les moments magnétiques différents ont tendance à se repousser, de sorte que si deux atomes voisins ont des spins différents, l’énergie du système est plus élevée que s’ils sont alignés — et pour la mesure de Gibbs (5.5), une énergie plus élevée donne une probabilité plus faible. On suppose en revanche que deux atomes qui ne sont pas immédiatement voisins n’interagissent pas directement l’un avec l’autre. On modélise ceci en posant

$$V(\sigma) = - \sum_{(a,b) | a \sim b} \sigma(a) \cdot \sigma(b)$$

où la somme porte sur l’ensemble des couples a, b de R qui sont voisins.

1. On considère la chaîne de Markov de matrice Q dont l’évolution est donnée comme suit : partant de σ , on choisit $a \in R$ uniformément et on renverse le spin en ce site, sans toucher au reste : on passe en σ' vérifiant $\sigma'(a) = -\sigma(a)$ et $\sigma'(b) = \sigma(b)$ pour $b \neq a$. Si l’on note Q la matrice de transition associée, a-t-elle la propriété $Q(\sigma, \sigma') > 0 \Rightarrow Q(\sigma', \sigma) > 0$? A-t-on une relation plus précise entre $Q(\sigma, \sigma')$ et $Q(\sigma', \sigma)$?
2. Supposons que dans l’évolution $\sigma \rightarrow \sigma'$ ci-dessus, on ait choisi de renverser le site a . Montrez que

$$\Delta V(\sigma, \sigma') := V(\sigma') - V(\sigma) = 2\sigma(a) \cdot \sum_{b | b \sim a} \sigma(b). \quad (5.6)$$

Définir une fonction qui calcule la quantité donnée en (5.6) si on lui donne σ (comme un array Numpy) et \mathbf{a} .

3. Reprendre l’algorithme de Metropolis–Hastings, version (5.3), pour Q et $\pi_{V,T}$. Observer que l’évolution est définie simplement comme suit : partant de σ ,
 - on choisit un site $a \in R$ uniformément, et on calcule $\Delta V(\sigma, \sigma')$;
 - si $\Delta V < 0$, on renverse le spin en a ;
 - si $\Delta V \geq 0$, on renverse le spin en a avec probabilité $e^{-\Delta V/T}$.
4. Montrer que la chaîne de Markov définie par l’algorithme de Metropolis–Hastings est irréductible apériodique. Ecrire une fonction qui exploite l’algorithme de Metropolis pour tirer une configuration de distribution proche de l’état de Gibbs du modèle ci-dessus.
5. Afficher de telles configuration pour différentes valeur de T , en utilisant la commande `matshow` de Matplotlib. On pourra prendre les valeurs $r = 20$, $n = 10000$ et $T = 9, 7, 5, 3, 1$.

6. Pour une configuration σ donnée, on peut définir l'aimantation macroscopique $\alpha(\sigma) = \frac{1}{r^2} \sum_{a \in R} \sigma(a)$. Pour différents tirages aux différentes valeurs de T données, calculez $|\alpha(\sigma)|$. La valeur observée indique-t-elle que les spins ont tendance à s'aligner, ou bien peut-elle être simplement l'effet du hasard ? Pour répondre à cette question, donnez un intervalle de fluctuation à 95% pour $|\alpha(\sigma)|$ dans le cas où σ est obtenu en tirant uniformément, en chaque site, un spin ± 1 .

5.5 Méthode du recuit simulé

Nous allons maintenant voir un autre intérêt des méthodes de simulation par chaînes de Markov. Le problème consiste à trouver un minimum de la fonction d'énergie V sur E . Dans certaines situations c'est très difficile : on traite dans l'exercice 62 un exemple où ce problème est NP-complet, avec le problème du *voyageur de commerce*. Notons

$$V_{\min} = \{x \in E \mid V(x) = \inf_E V\}.$$

5.5.1 Algorithme du recuit

Lemme 5.5.1 *On a la convergence*

$$\pi_{V,0} \stackrel{\text{def.}}{=} \lim_{T \rightarrow 0} \pi_{V,T} = \frac{1}{\text{card } V_{\min}} \sum_{x \in V_{\min}} \delta_x.$$

Preuve. Soient x, y deux points de E . On a

$$\frac{\pi_{V,T}(y)}{\pi_{V,T}(x)} = e^{-(V(y)-V(x))/T}.$$

Par conséquent, si $V(y) > V(x)$ alors $\lim_{T \rightarrow 0} \frac{\pi_{V,T}(y)}{\pi_{V,T}(x)} = 0$ donc $\lim_{T \rightarrow 0} \pi_{V,T}(y) = 0$. Cela est nécessairement vrai pour tout $y \notin V_{\min}$. Pour $x, y \in V_{\min}$ on a $\pi_{V,T}(x) = \pi_{V,T}(y)$ pour tout T . \square

On voudrait donc de simuler par l'algorithme de Metropolis-Hastings les mesures $\pi_{V,T}$ pour des T de plus en plus petits. On note P_T la matrice de transition associée. Le problème est que le temps de convergence de la chaîne vers sa mesure d'équilibre est typiquement d'ordre $\exp\left(\frac{C}{T}\right)$ (on donne quelques éléments de justification dans la partie 5.5.2). L'idée est donc de diminuer T par paliers, en attendant à chaque fois un peu plus longtemps.

Plus précisément (voir Théorème 3.3.11 et section 3.3.4 de *Promenade aléatoire* de Benaïm et El Karoui), on considère une chaîne de Markov obtenue par l'algorithme de Metropolis à partir d'une mesure de Gibbs associée à un potentiel dont on suppose qu'il vérifie

$$Q(x, y) > 0 \Rightarrow V(x) \neq V(y). \quad (5.7)$$

Théorème 5.5.2 *On suppose que V a la propriété (5.7) et que $\inf_E V > 0$. Il existe une constante C ne dépendant que de V telle que si l'on choisit une suite de températures $(T(n))_n$ données par*

$$T(n) = \frac{1}{k} \quad \text{pour} \quad e^{C(k-1)} \leq n < e^{Ck},$$

alors on a

$$\lim_n P_{T(n)} \dots P_{T(1)} f = \frac{1}{\text{card } V_{\min}} \sum_{x \in V_{\min}} f(x).$$

On obtient donc sous les hypothèses du Théorème, que partant de n'importe quel $x \in E$ on atteindra presque-sûrement un minimum de V . Cet algorithme s'appelle le "recuit simulé", par analogie avec la technique métallurgique où l'on obtient un métal durci en le chauffant avant de le laisser refroidir lentement, et ce plusieurs fois.

Une description rapide de cet algorithme est la suivante : on teste des changements de configuration en les sélectionnant au hasard suivant Q . Si le changement fait baisser V (cas $\Delta V < 0$) alors on accepte la nouvelle configuration. Si le changement fait augmenter V (cas $\Delta V > 0$), on peut l'accepter ou le refuser, suivant que $U \leq e^{-\Delta V/T}$ ou non. Le premier mécanisme tend à faire diminuer V ; le deuxième évite que l'on se retrouve coincé en un minimum local. Le paramètre modifiant la tendance à accepter un changement défavorable est T (on l'accepte d'autant plus que T est grand) ; choisir T décroissant vers 0 nous assure que l'on finira par se fixer en un minimum, la décroissance lente doit nous assurer que l'on aura pris assez de risque pour explorer "toutes" les possibilités avant de se fixer.

Les obstructions techniques sont de deux types :

- on doit choisir un schéma de décroissance par palier dépendant d'une constante C , mais ce C est difficile à calculer en pratique),
- on a la convergence presque-sûre, mais expliciter la vitesse de convergence est difficile.

En pratique, il sera facile de voir si l'algorithme converge vers un minimum de V . En faisant tourner des simulations, on verra si l'algorithme se comporte bien, et en particulier s'il semble avoir l'une des pathologies suivantes :

- une convergence trop lente, due à un algorithme qui continue trop longtemps à accepter les sélections obtenues par Q (ce qui se produit quand la température décroît trop lentement) ;
- une convergence trop rapide, en général vers un minimum local, due à un algorithme qui se met trop rapidement à refuser les sélections obtenues par Q (ce qui se produit quand la température décroît trop rapidement).

On aura donc intérêt à jouer à varier la valeur de C ... ou même à changer de schéma de décroissance par palier.

Exercice 62 Un voyageur doit visiter r villes, que l'on représente par r points M_1, \dots, M_r du plan. Sa ville de départ doit être la même que sa ville d'arrivée, mais on suppose (cela simplifie les notations) qu'il peut choisir cette ville aussi. Puisqu'il passera dans chaque ville une et une seule fois, son parcours est déterminé par une permutation σ de $\{1, \dots, r\}$, donc $E = \mathfrak{S}_r$. Les villes visitées sont alors, dans l'ordre, $M_{\sigma(1)}, M_{\sigma(2)} \dots, M_{\sigma(r)}$. La distance entre les villes i et j sont données par $d(i, j)$; le voyageur parcourra donc une distance

$$V(\sigma) = \sum_{i=1}^r d(M_{\sigma(i)}, M_{\sigma(i+1)}) \quad (5.8)$$

où l'on considère que $r+1 = 1$. On souhaite trouver un itinéraire σ pour lequel la distance parcourue totale est minimale. Il y a $r!$ itinéraires; tous les tester serait beaucoup trop long. On va donc utiliser le recuit simulé.

1. Ecrire une fonction Python qui, pour des variables d'entrée `M` et `sigma` qui sont respectivement une liste de paires de points représentant les coordonnées $(x_1, y_1), \dots, (x_r, y_r)$ de r points M_1, \dots, M_r du plan et une permutation de $\{1, \dots, n\}$, calcule la distance totale $V(\sigma)$ telle que définie par la relation (5.8).
2. On considère la transition Q dont l'évolution est la suivante : partant de σ , on choisit au hasard et uniformément deux points distincts, qui s'écrivent donc $\sigma(i)$ et $\sigma(j)$. Si par exemple on visitait $\sigma(i)$ avant $\sigma(j)$ (c'est-à-dire si $i < j$) alors on échange $\sigma(i)$ et $\sigma(j)$ dans l'itinéraire (par exemple si le couple choisi est 3,6 alors $[2,3,1,4,6,5]$ devient $[2,6,1,4,3,5]$). Montrez que l'évolution en question a la propriété $Q(\sigma, \sigma') = Q(\sigma', \sigma)$, puis écrivez une fonction qui pour une variable d'entrée `sigma` représentant une permutation σ , retourne `sigmap` représentant σ' définie comme ci-dessus.
3. Pour un $T > 0$, reprenez l'algorithme de Metropolis-Hastings pour Q et $\pi_{V,T}$. Observez que l'évolution de matrice de transition P_T est définie simplement comme suit : partant de σ ,
 - (a) on modifie l'itinéraire σ en σ' comme ci-dessus,
 - (b) on calcule $\Delta V = V(\sigma') - V(\sigma)$ et on simule une uniforme $U \sim \mathcal{U}([0,1])$,
 - (c) si $U < e^{-\Delta V/T}$, on sélectionne σ' , sinon on conserve σ .
4. Pour les points M_1, \dots, M_r , on tire r points au hasard³ dans $[0,1]^2$. Appliquer l'évolution ci-dessus avec une température variable suivant le schéma donné dans le Théorème 5.5.2, en faisant varier la valeur de C . On pourra essayer aussi avec une variation plus rapide $T(n) = C/n$ et comparer les résultats.

3. Si l'on est un peu plus curieux de vraies données, on peut aller récupérer les positions des villes françaises sur data.gouv.fr

5.5.2 Vitesse de convergence : méthode spectrale

Donnons ici quelques éléments sur la vitesse de convergence de la chaîne P_T , qui sont les résultats sous-jacents au Théorème 5.5.2. L'un des intérêts mathématiques de l'hypothèse de réversibilité vient du résultat suivant :

Lemme 5.5.3 *La matrice P est réversible par rapport à π si et seulement si elle est autoadjointe pour le produit scalaire défini par $\langle f, g \rangle_\pi = \sum_{x \in E} f(x)g(x)\pi(x)$. Dans ce cas, ses valeurs propres peuvent s'écrire*

$$1 \geq v_1 \geq v_2 \geq \dots \geq v_d \geq -1.$$

Si de plus P est irréductible, alors $1 = v_1 > v_2$. Si de plus P est irréductible et apériodique alors $v_d > -1$.

Sous l'hypothèse que P est irréductible et apériodique alors $P^n f$ va tendre vers la projection de f sur le vecteur propre de P associé à 1. Comme celui-ci est la fonction constante égale à 1 et que la projection est par rapport à $\langle \cdot, \cdot \rangle_\pi$ on a $\xrightarrow{n \rightarrow \infty} \int f d\pi$. On peut améliorer ce résultat et obtenir une vitesse de convergence en notant $\rho = \max(|v_2|, |v_d|)$. Pour toute fonction f sur E on a pour tout n :

$$\|P^n f - \int f d\pi\|_\pi \leq \rho^n \|f - \int f d\pi\|_\pi.$$

En pratique, des calculs plus pénibles que réellement compliqués (voir encore *Promenade aléatoire* de Benaïm et El Karoui) permettent de montrer, pour une chaîne de Markov obtenue par l'algorithme de Metropolis à partir d'une mesure de Gibbs associée à un potentiel qui vérifie (5.7), que l'on a une inégalité

$$\rho \leq 1 - d^{-3} e^{-C(V)/T}$$

où $d = \text{card } E$ et $C(V)$ est la fameuse constante qui ne dépend que de V (et en particulier pas de T).

Ainsi, partant de n'importe quelle distribution μ , la probabilité μ_n converge exponentiellement vite vers $\pi_{V,T}$ – mais avec un taux

$$-\log\left(1 - \frac{1}{d^3} e^{-C(V)/T}\right) \underset{T \rightarrow 0}{\sim} d^{-3} e^{-C(V)/T}$$

donc qui s'approche rapidement de zéro lorsque $T \rightarrow 0$.

5.6 Exercice supplémentaires

Exercice 63 *On place un cavalier sur un échiquier, sur la case A1. À chaque étape, on le déplace sur une des cases accessibles (par son déplacement habituel) choisie uniformément. Quelle est l'espérance de son temps de retour en A1 ?*

Exercice 64 *On considère un sous-graphe (G, A) du réseau $\{1, \dots, 10\}^3$, donné par un ensemble de points G et une matrice d'adjacence symétrique A . On veut "ranger" dans cette boîte des boules de diamètre R vérifiant $1 < R < \sqrt{2}$,*

de sorte que l'on ne peut mettre deux boules sur des sites voisins mais qu'il n'y a pas d'autre contrainte. On veut utiliser le recuit simulé pour trouver une configuration permettant de ranger le plus grand nombre possible de boules dans la boîte. On décrit par $x : \Lambda \rightarrow \{0,1\}$ une configuration.

1. Montrez que l'algorithme suivant : partant de x ,
 - (a) on choisit un site s au hasard dans G ,
 - (b) si tous les sites voisins (au sens de l'adjacence dans S) sont libres, on ajoute une boule en s ,définit une matrice de transition Q vérifiant $Q(x,y) = 0 \Leftrightarrow Q(y,x) = 0$.
2. Montrez que la fonction V définie par

$$V(x) = 1000 - \text{card}\{s \in \Lambda \mid x(s) = 1\}$$

(donc le nombre de sites libres dans la configuration x) vérifie bien pour $x \neq y$ la relation $Q(x,y) > 0 \Rightarrow V(x) \neq V(y)$.

3. Appliquez l'algorithme du recuit simulé pour trouver une configuration avec un nombre maximal de boules. On pourra faire varier le schéma de décroissance de la température en fonction du comportement observé de l'algorithme. Comme exemples de (G,A) on pourra partir du réseau cubique $\{1, \dots, 10\}^3$ et lui enlever aléatoirement des arêtes et/ou sommets.